Machine Learning Department
School of Computer Science
Carnegie Mellon University

Final Exam Review

Matt Gormley
Lecture 29
Apr. 29, 2020

Reminders

- Homework 9: Learning Paradigms
  - Out: Wed, Apr. 22
  - Due: Wed, Apr. 29 at 11:59pm
  - Can only be submitted up to 3 days late, so we can return grades before final exam
- Final Exam Practice Problems
  - Out: Wed, Apr. 29
- Final Exam
  - Mon, May 04 (1pm – 4pm)

- Today's In-Class Poll
  - http://poll.mlcourse.org

EXAM LOGISTICS

Final Exam

- **Time / Location**
  - **Time:** Registrar-scheduled Exam
    Mon, May 4th at 1:00pm – 4:00pm
  - **Online Exam:** Same format as Midterm Exam 2
  - Please watch Piazza carefully for announcements logistics
- **Logistics**
  - Distribution of Topics: Lectures 19 – 28 (95%), Lectures 1 – 18 (5%)
  - Format of questions:
    - Multiple choice
    - True / False (with justification)
    - Derivations
    - Short answers
    - Interpreting figures
    - Implementing algorithms on paper
  - You are encouraged to **bring** one 8½ x 11 sheet of notes (front and back)
  - Open book according to my definition on Piazza:
    https://piazza.com/class/k4wzus8w2c11u6?cid=1673

Final Exam

- **How to Prepare**
  - Attend (or watch) this final exam review session
  - Review Practice Problems: Exam 3
    - Disclaimer: the practice problems are somewhere between homework-style problems and exam-style problems
  - Review this year's **homework problems**
  - Review the **poll questions** from each lecture
  - Consider whether you have achieved the **learning objectives** for each lecture / section

Final Exam

- **Advice (for during the exam)**
  - Solve the easy problems first
    (e.g. multiple choice before derivations)
    - if a problem seems extremely complicated you're likely missing something
  - Don't leave any answer blank!
  - If you make an assumption, write it down
  - If you look at a question and don't know the answer:
    - we probably haven't told you the answer
    - but we've told you enough to work it out
    - imagine arguing for some answer and see if you like it

Topics for Midterm 1

- Foundations
  - Probability, Linear Algebra, Geometry, Calculus
  - Optimization
- Important Concepts
  - Overfitting
  - Experimental Design
- Classification
  - Decision Tree
  - KNN
  - Perceptron
- Regression
  - Linear Regression

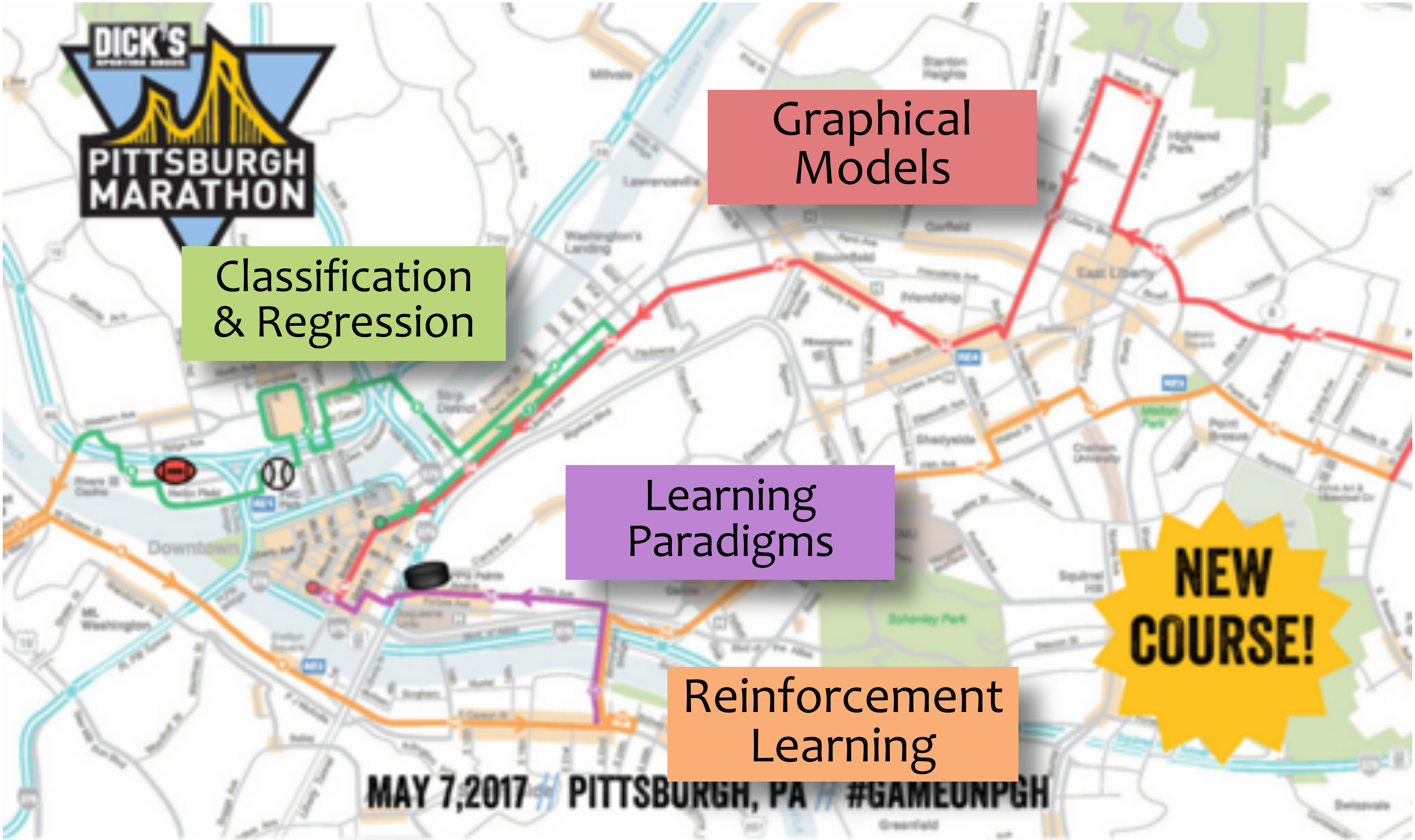
Topics for Midterm 2

- Classification
  - Binary Logistic Regression
  - Multinomial Logistic Regression
- Important Concepts
  - Stochastic Gradient Descent
  - Regularization
  - Feature Engineering
- Feature Learning
  - Neural Networks
  - Basic NN Architectures
  - Backpropagation
- Learning Theory
  - PAC Learning
- Generative Models
  - Generative vs. Discriminative
  - MLE / MAP
  - Naïve Bayes

Topics for Final Exam

- Graphical Models
  - HMMs
  - Learning and Inference
  - Bayesian Networks
- Reinforcement Learning
  - Value Iteration
  - Policy Iteration
  - Q-Learning
  - Deep Q-Learning
- Other Learning Paradigms
  - K-Means
  - PCA
  - SVM (large-margin)
  - Kernels
  - Ensemble Methods
  - Recommender Systems


DICK'S
SPORTING GOODS
PITTSBURGH
MARATHON


DICK'S
PITTSBURGH
MARATHON
Graphical
Models
Classification
& Regression
Learning
Paradigms
Reinforcement
Learning
NEW
COURSE!
Downtown
MAY 7,2017 // PITTSBURGH, PA // #GAMEONPGH

RICHARD S. CALIGUIRI
CITY OF PITTSBURGH
Great Race

Learning as Memorization

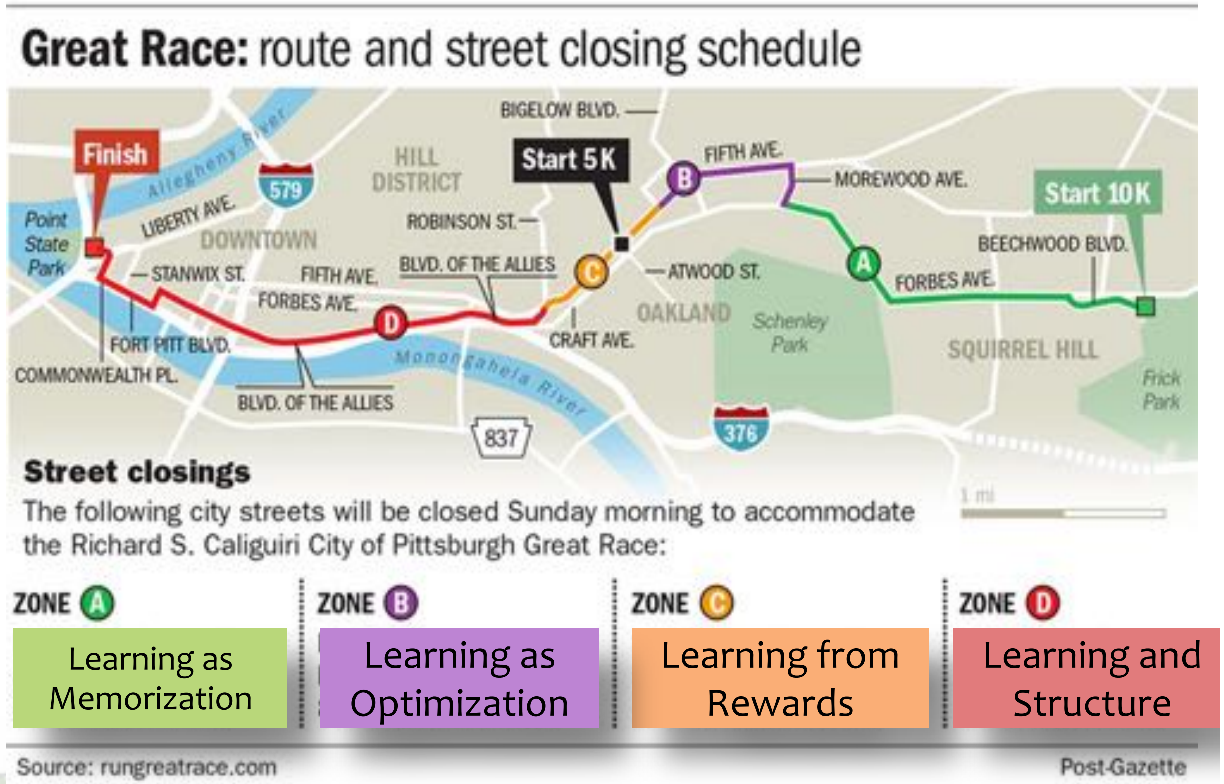
Learning as Optimization

Learning from Rewards

Learning and Structure

A new **combined** course…

… with the best (uphill climbs) from both



Material Covered **Before** Midterm Exam 2

SAMPLE QUESTIONS

Matching Game

Goal: Match the Algorithm to its Update Rule

1. SGD for Logistic Regression $h_{\boldsymbol{\theta}}(\mathbf{x}) = p(y\|x)$	**4.** $\theta_k \leftarrow \theta_k + (h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)})$
2. Least Mean Squares $h_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x}$	**5.** $\theta_k \leftarrow \theta_k + \frac{1}{1 + \exp \lambda (h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)})}$
3. Perceptron (next lecture) $h_{\boldsymbol{\theta}}(\mathbf{x}) = \text{sign}(\boldsymbol{\theta}^T \mathbf{x})$	**6.** $\theta_k \leftarrow \theta_k + \lambda (h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)}) x_k^{(i)}$

A. 1=5, 2=4, 3=6
B. 1=5, 2=6, 3=4
C. 1=6, 2=4, 3=4
D. 1=5, 2=6, 3=6
E. 1=6, 2=6, 3=6

Use Probability

Sample Questions

1.4 Probability

Assume we have a sample space Ω. Answer each question with **T** or **F**.

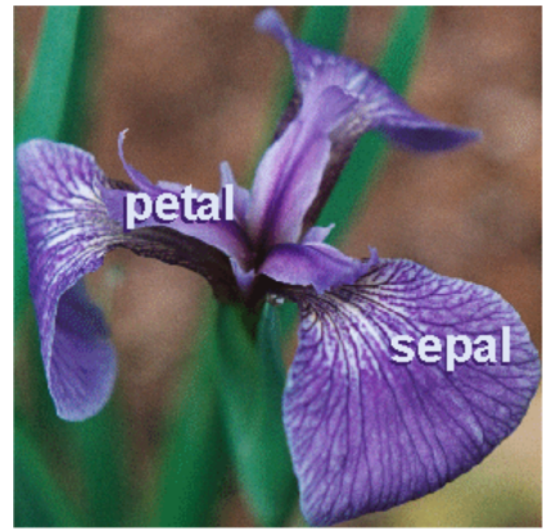
(a) [1 pts.] **T or F:** If events A, B, and C are disjoint then they are independent.

(b) [1 pts.] **T or F:** $P(A|B) \propto \frac{P(A)P(B|A)}{P(A|B)}$. (The sign '$\propto$' means 'is proportional to')

Medical Diagnosis

Interview Transcript
Date: Jan. 15, 2020.
Parties: Matt Gormley and Doctor E.
Topic: Medical decision making

- Matt: Welcome. Thanks for interviewing with me today.
- Dr. E: Interviewing…?
- Matt: Yes. For the record, what type of doctor are you?
- Dr. E: Who said I'm a doctor?
- Matt: I thought when we set up this interview you said—
- Dr. E: I'm a preschooler.
- Matt: Good enough. Today, I'd like to learn how you would determine whether or not your little brother is sick given his symptoms.
- Dr. E: He's not sick.
- Matt: We haven't started yet. Now, suppose he is sneezing. Is he sick?
- Dr. E: No, that's just the sniffles.
- Matt: What if he is coughing; Is he sick?
- Dr. E: No, he just has a cough.
- [Editor's note: preschoolers unilaterally agree that having the sniffles or a cough is not the same as being sick.]
- Matt: What if he's both sneezing and coughing?
- Dr. E: Then he's sick.
- Matt: Got it. What if your little brother is sneezing and coughing, plus he's a doctor.
- Dr. E: Then he's not sick.
- Matt: How do you know?
- Dr. E: Doctors don't get sick.
- Matt: What if he is not sneezing, but is coughing, and he is a fox….
- Matt: …and the fox is in the bottle where the tweetle beetles battle with their paddles in a puddle on a noodle-eating poodle.
- Dr. E: Then he is must be a tweetle beetle noodle poodle bottled paddled muddled duddled fuddled wuddled fox in socks, sir. That means he's definitely sick.
- Matt: Got it. Can I use this conversation in my lecture?
- Dr. E: Yes



Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
1	6.7	3.0	5.0	1.7

Sample Questions

4 K-NN [12 pts]

Now we will apply K-Nearest Neighbors using Euclidean distance to a binary classification task. We assign the class of the test point to be the class of the majority of the k nearest neighbors. A point can be its own neighbor.

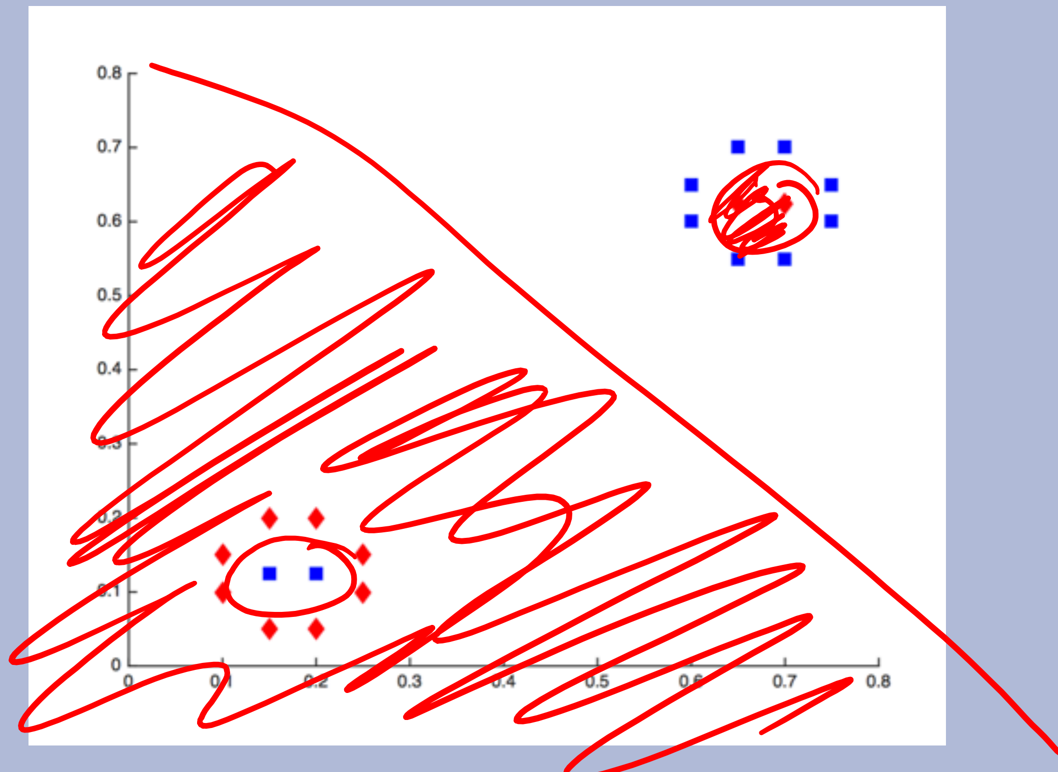
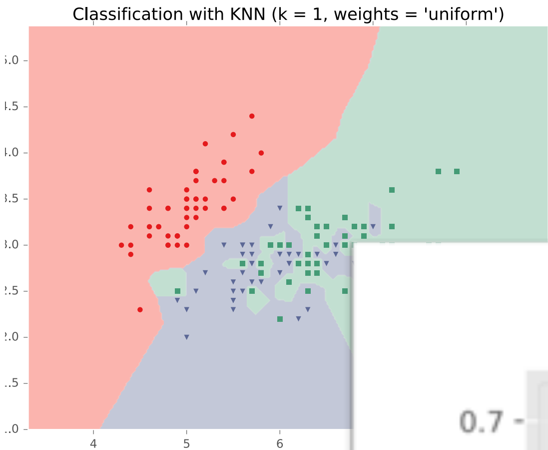
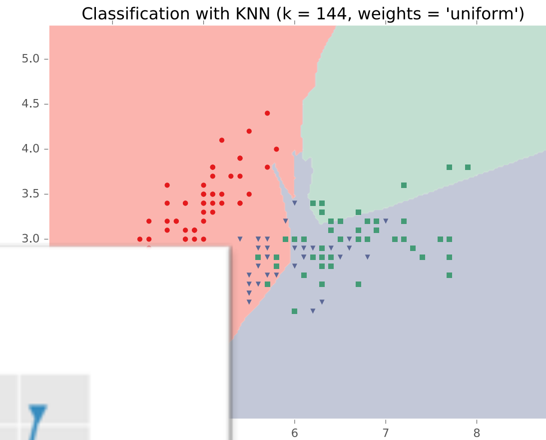


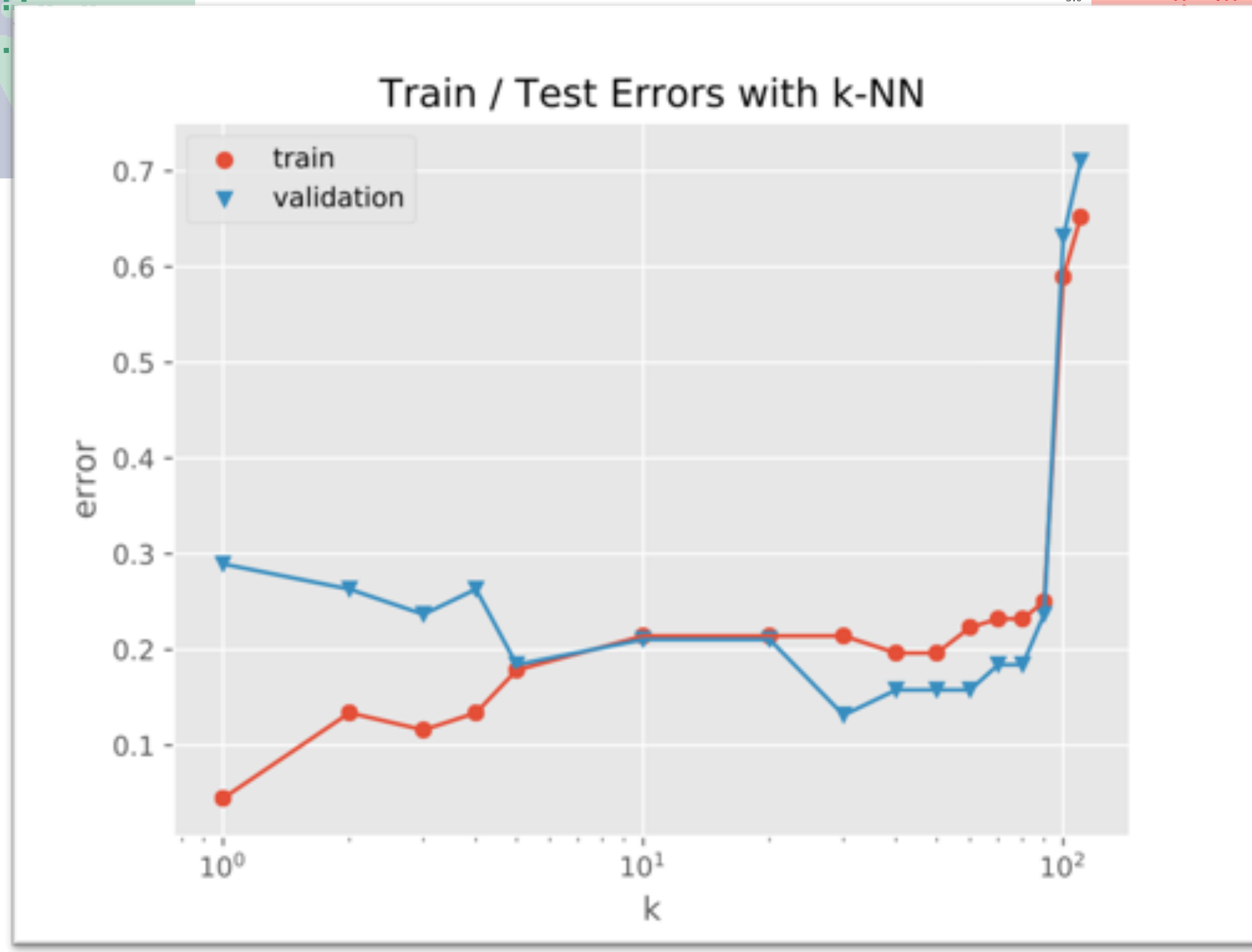
Figure 5

3. **[2 pts]** What value of k minimizes leave-one-out cross-validation error for the dataset shown in Figure 5? What is the resulting error?

k-NN: Choosing k

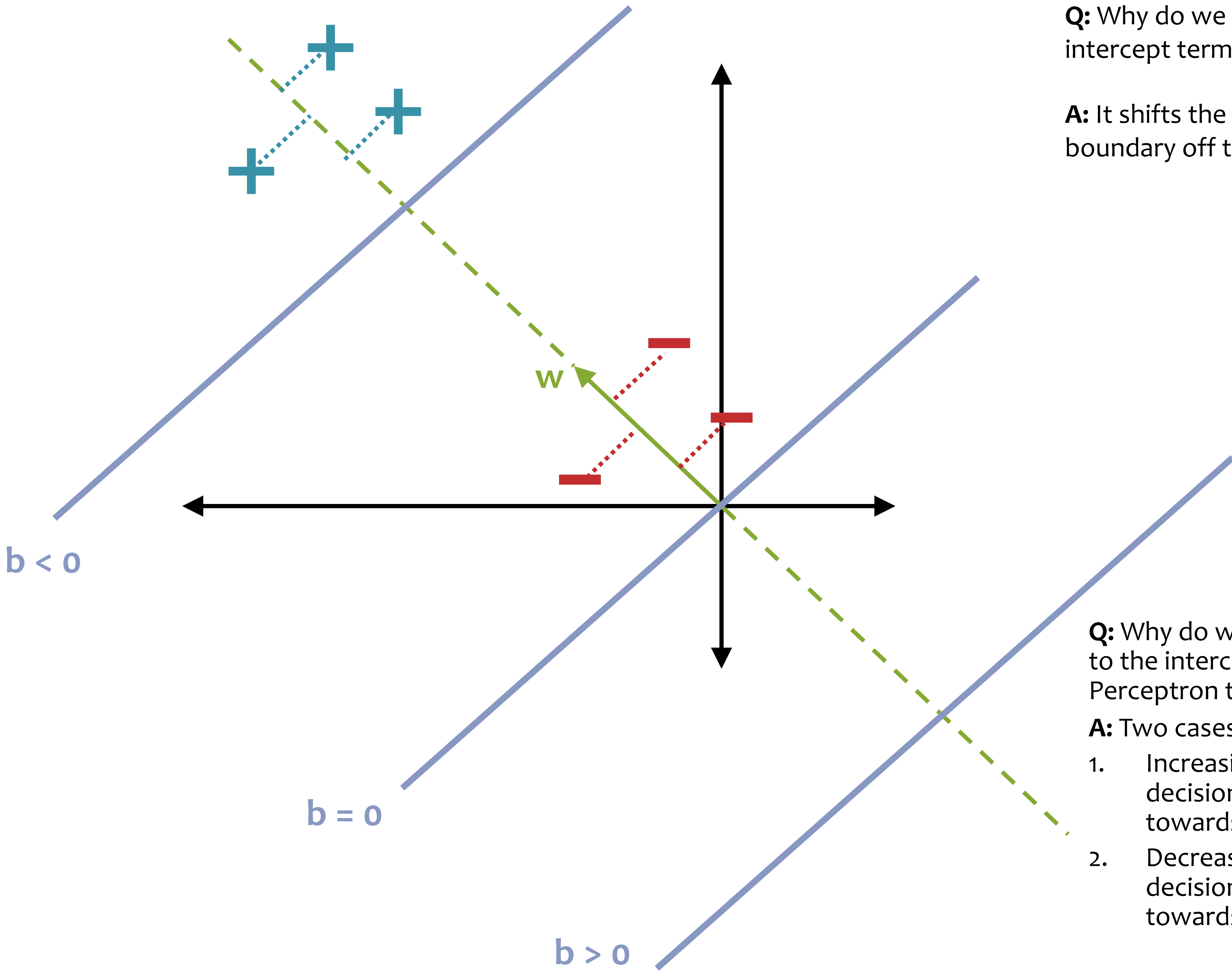






Fisher Iris Data: varying the value of k

Perceptron & The Intercept Term



Q: Why do we need an intercept term?

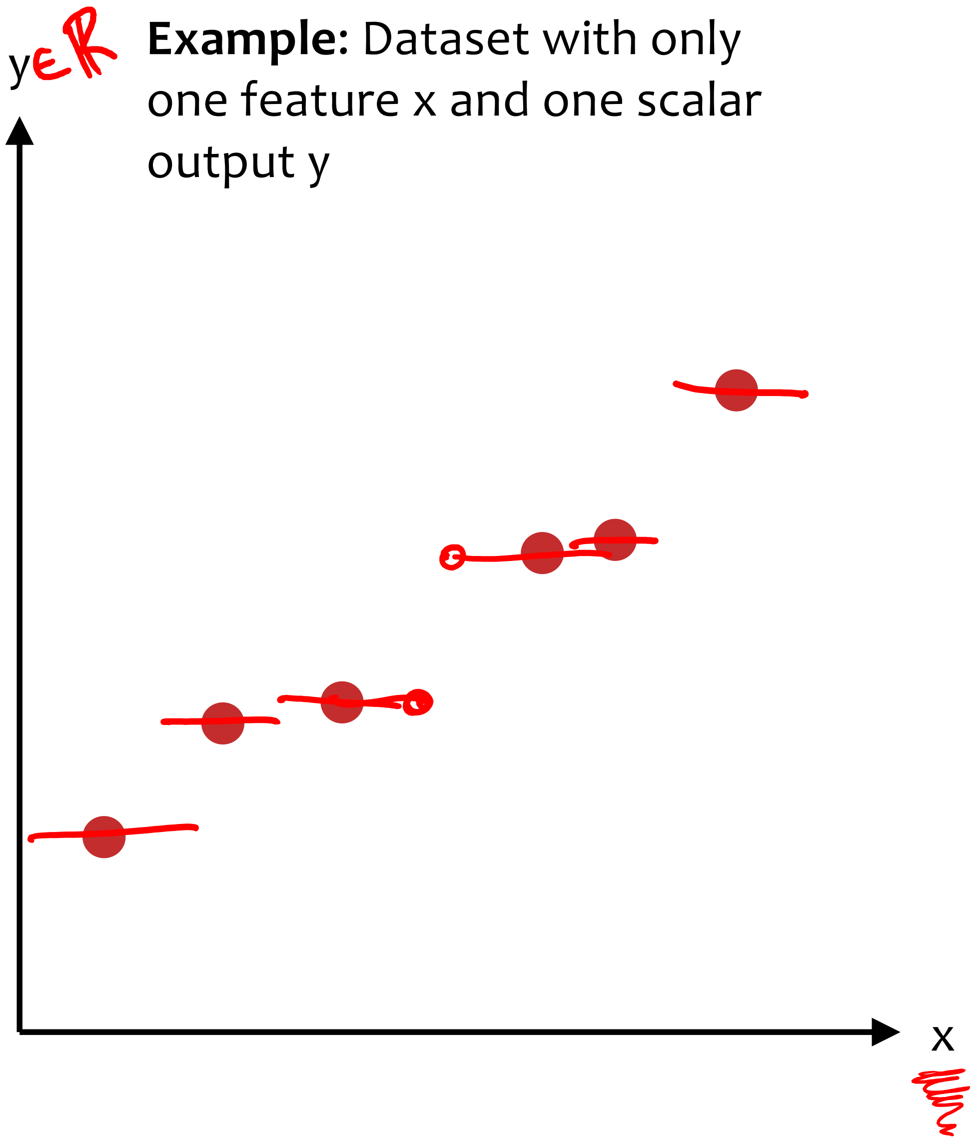
A: It shifts the decision boundary off the origin

Q: Why do we add / subtract 1.0 to the intercept term during Perceptron training?

A: Two cases

1. Increasing b shifts the decision boundary towards the negative side
2. Decreasing b shifts the decision boundary towards the positive side

k-NN Regression



Example: Dataset with only one feature x and one scalar output y

k=1 Nearest Neighbor Regression

- *Train*: store all (x, y) pairs
- *Predict*: pick the nearest x in training data and return its y

k=2 Nearest Neighbor Distance Weighted Regression

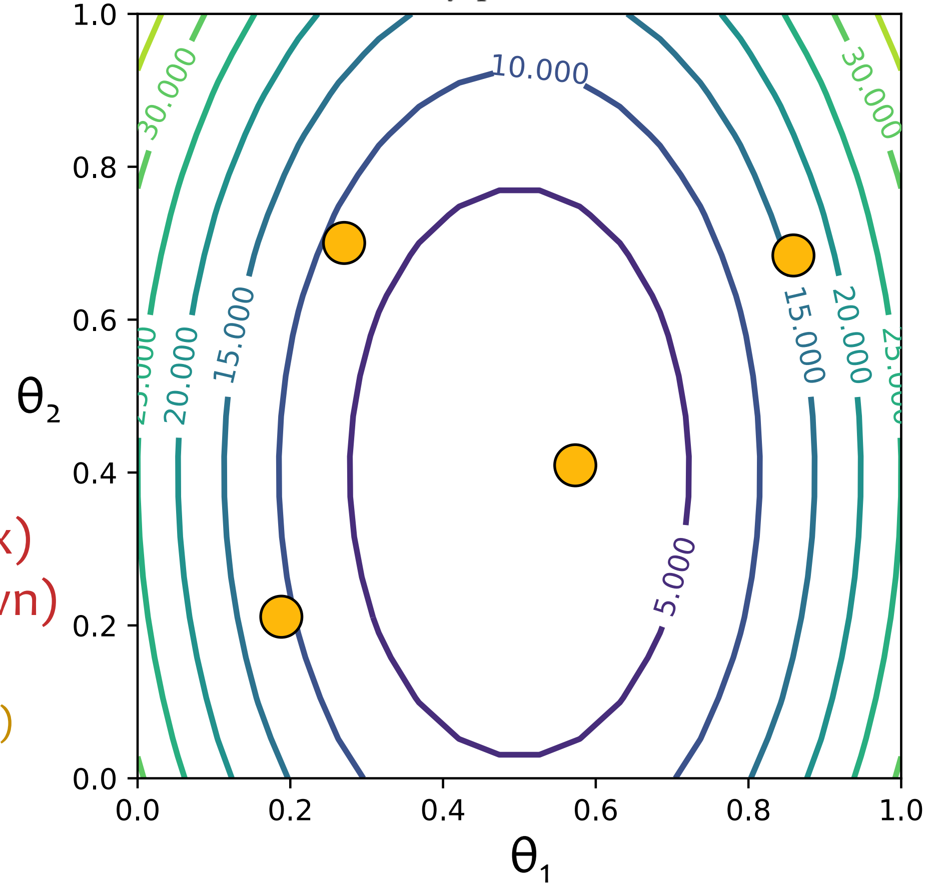
- *Train*: store all (x, y) pairs
- *Predict*: pick the nearest two instances $x^{(n1)}$ and $x^{(n2)}$ in training data and return the weighted average of their y values

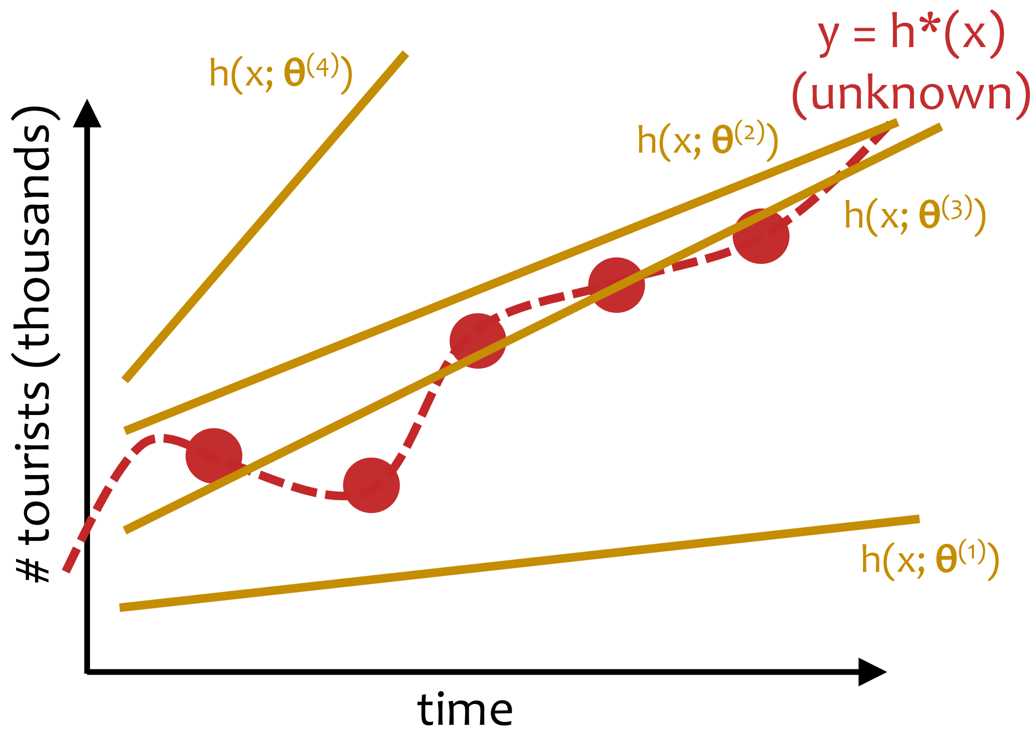
Linear Regression by Rand. Guessing

Optimization Method #0: Random Guessing

1. Pick a random $\boldsymbol{\theta}$
2. Evaluate $J(\boldsymbol{\theta})$
3. Repeat steps 1 and 2 many times
4. Return $\boldsymbol{\theta}$ that gives smallest $J(\boldsymbol{\theta})$

$$J(\boldsymbol{\theta}) = J(\theta_1, \theta_2) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \boldsymbol{\theta}^T\mathbf{x}^{(i)}\right)^2$$





t	θ_1	θ_2	$J(\theta_1, \theta_2)$
1	0.2	0.2	10.4
2	0.3	0.7	7.2
3	0.6	0.4	1.0
4	0.9	0.7	19.2

Sample Questions

3.1 Linear regression

Consider the dataset S plotted in Fig. 1 along with its associated regression line. For each of the altered data sets S^{new} plotted in Fig. 3, indicate which regression line (relative to the original one) in Fig. 2 corresponds to the regression line for the new data set. Write your answers in the table below.

Dataset	(a)	(b)	(c)	(d)	(e)
Regression line					

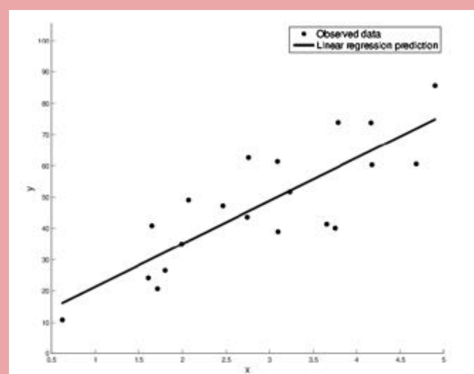


Figure 1: An observed data set and its associated regression line.

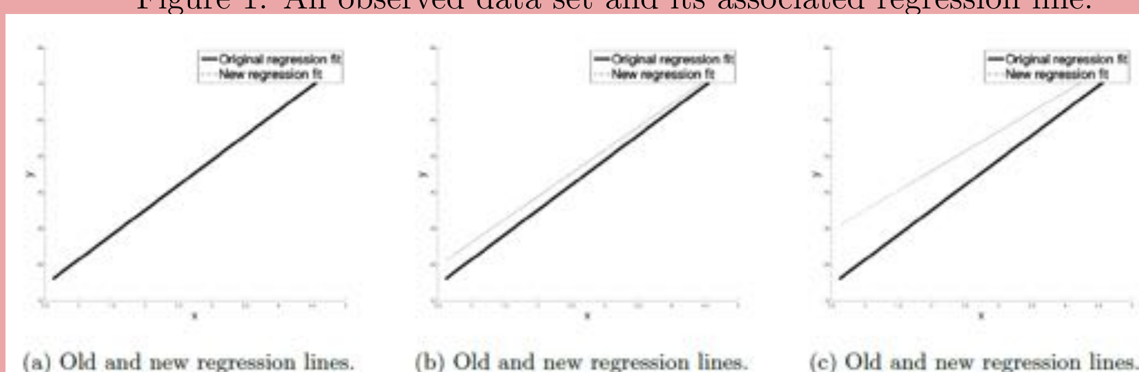
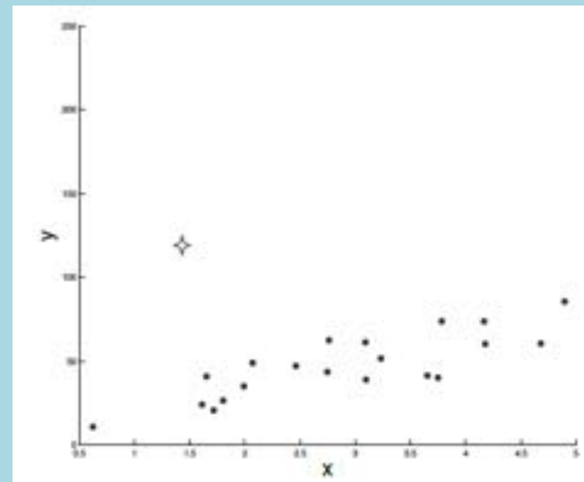


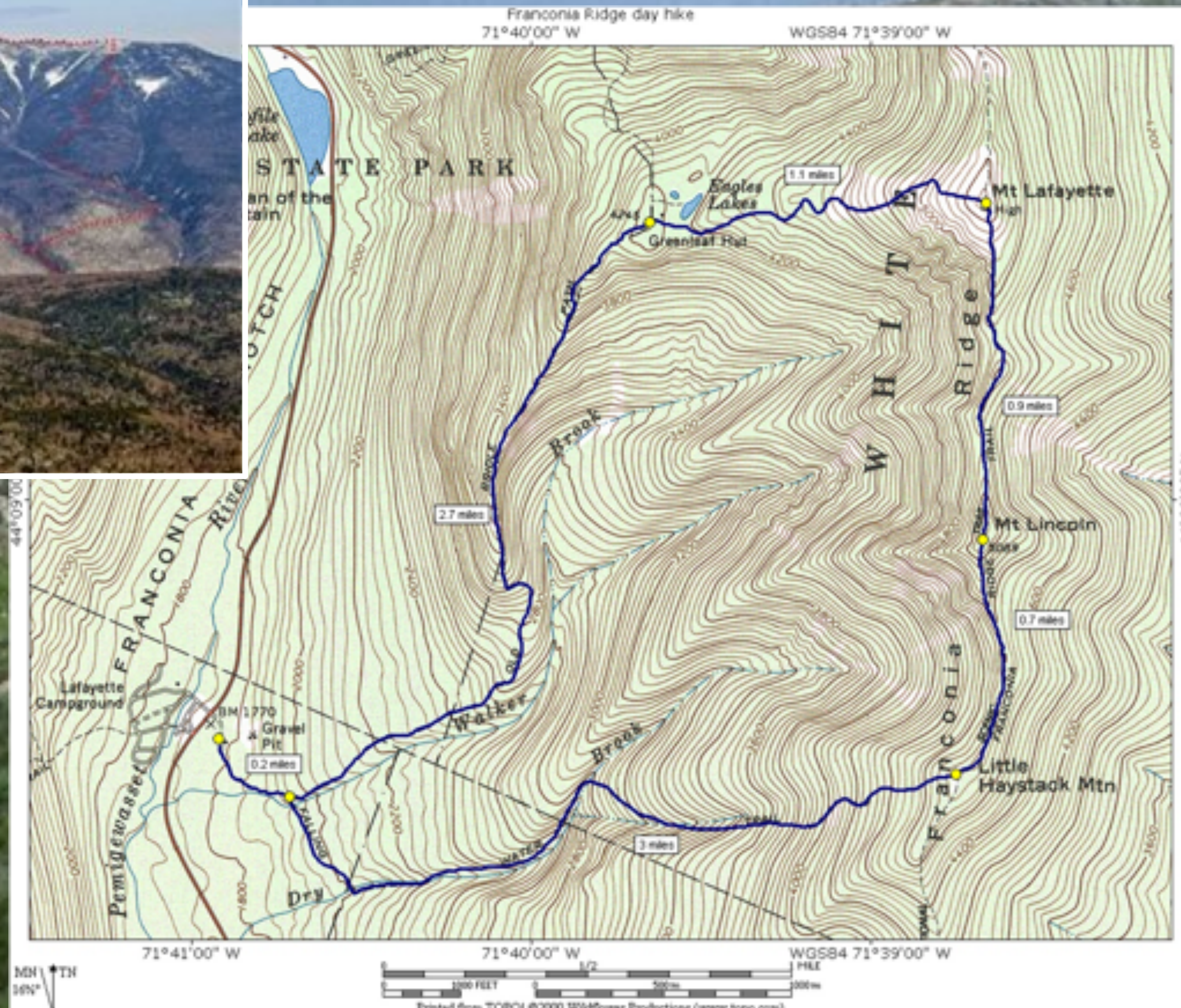
Figure 2: New regression lines for altered data sets S^{new}.

Dataset

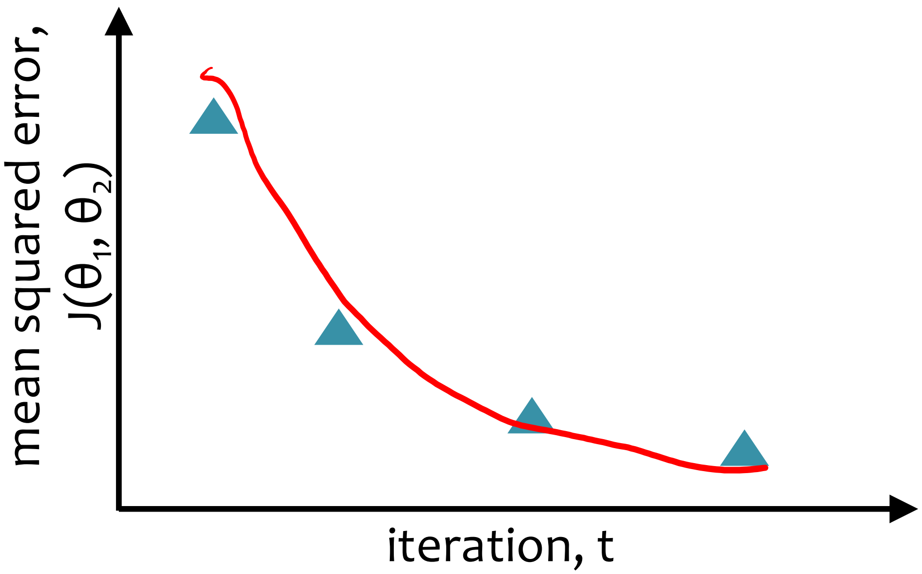


(a) Adding one outlier to the original data set.

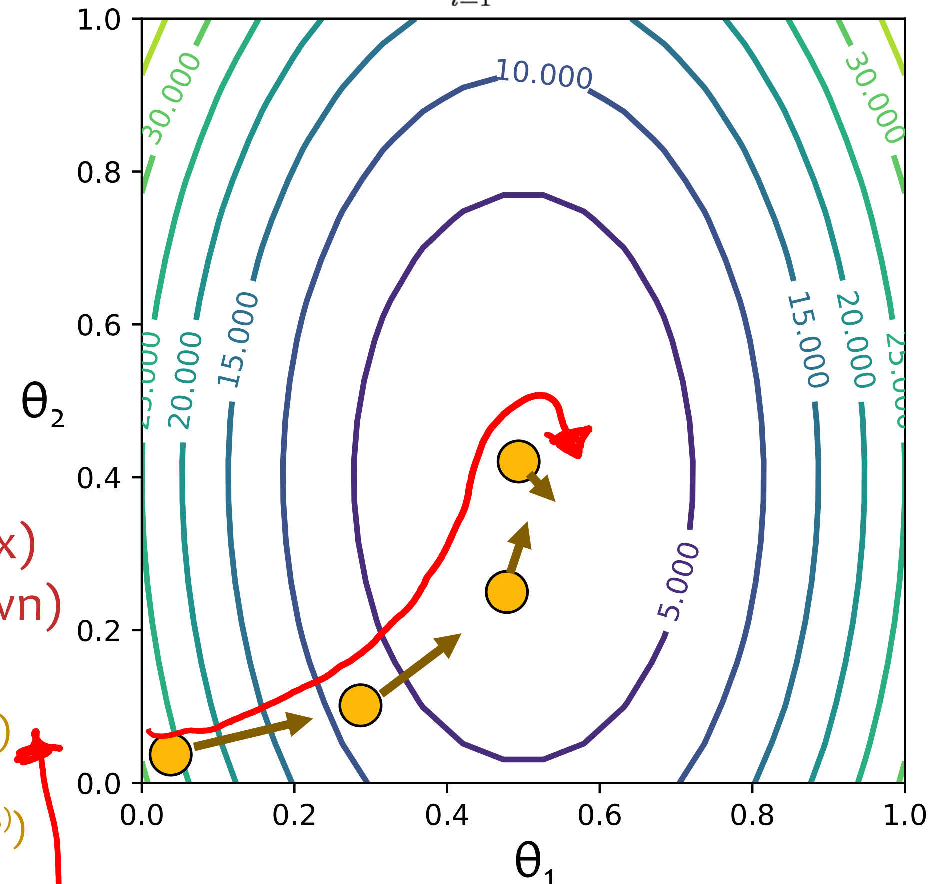
Topographical Maps

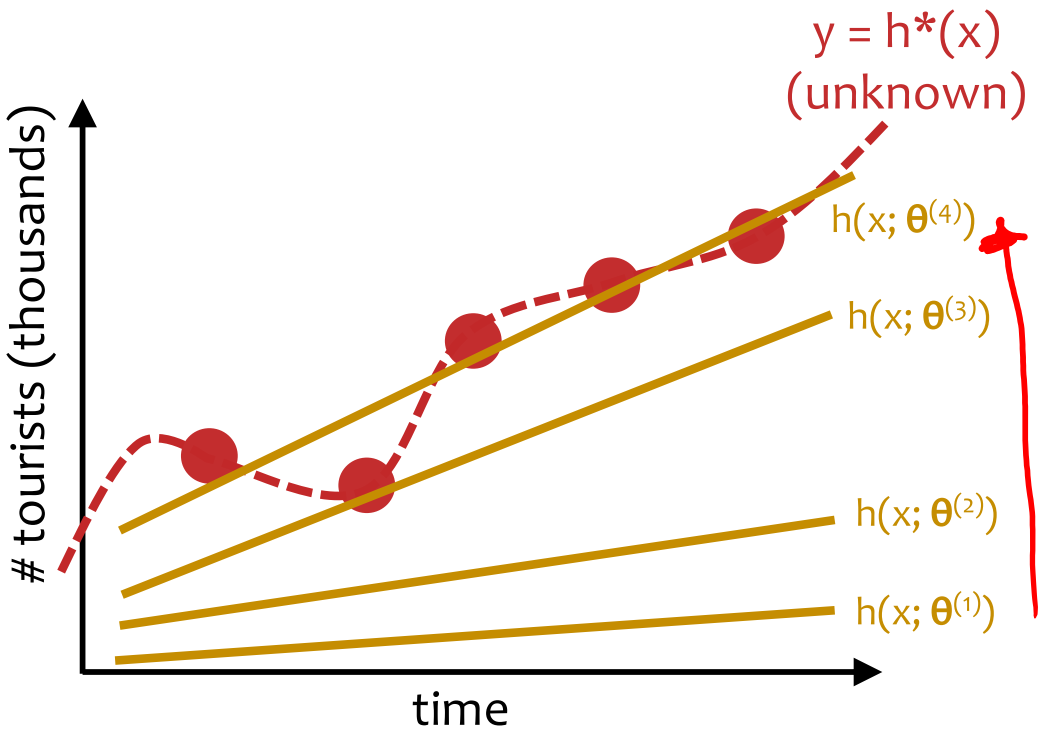


Linear Regression by Gradient Desc.



$$J(\boldsymbol{\theta}) = J(\theta_1, \theta_2) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \boldsymbol{\theta}^T \mathbf{x}^{(i)}\right)^2$$





t	θ_1	θ_2	$J(\theta_1, \theta_2)$
1	0.01	0.02	25.2
2	0.30	0.12	8.7
3	0.51	0.30	1.5
4	0.59	0.43	0.2

Sample Questions

3.1 Linear regression

Consider the dataset S plotted in Fig. 1 along with its associated regression line. For each of the altered data sets S^{new} plotted in Fig. 3, indicate which regression line (relative to the original one) in Fig. 2 corresponds to the regression line for the new data set. Write your answers in the table below.

Dataset	(a)	(b)	(c)	(d)	(e)
Regression line					

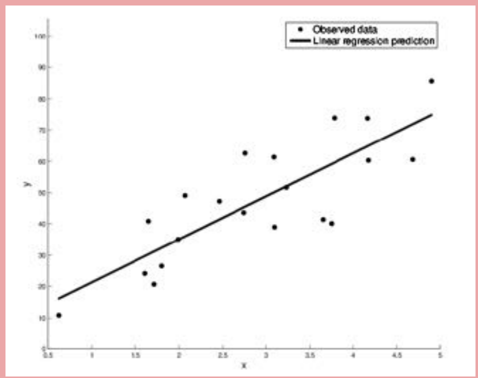
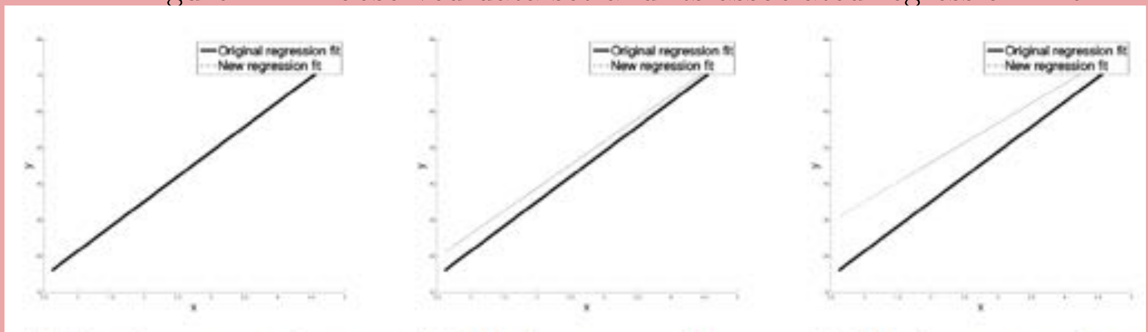


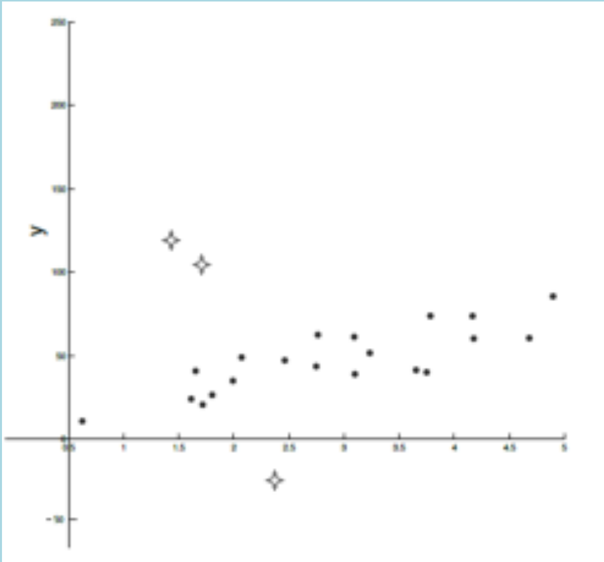
Figure 1: An observed data set and its associated regression line.



(a) Old and new regression lines. (b) Old and new regression lines. (c) Old and new regression lines.

Figure 2: New regression lines for altered data sets S^{new}.

Dataset



(c) Adding three outliers to the original data set. Two on one side and one on the other side.

Sample Questions

3.1 Linear regression

Consider the dataset S plotted in Fig. 1 along with its associated regression line. For each of the altered data sets S^{new} plotted in Fig. 3, indicate which regression line (relative to the original one) in Fig. 2 corresponds to the regression line for the new data set. Write your answers in the table below.

Dataset	(a)	(b)	(c)	(d)	(e)
Regression line					

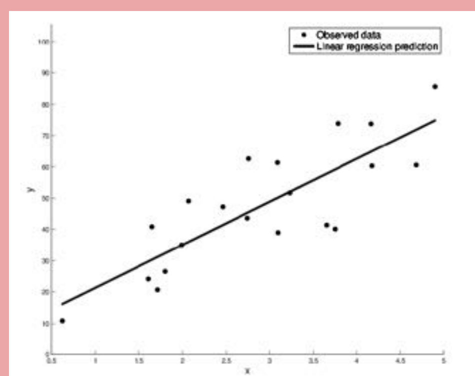
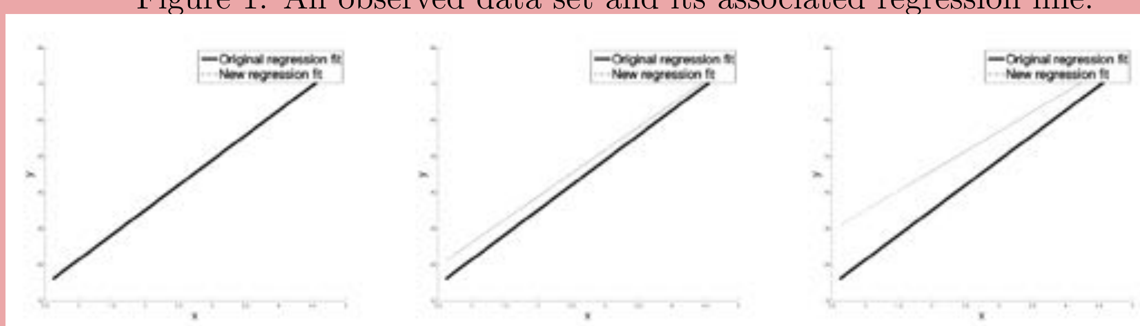


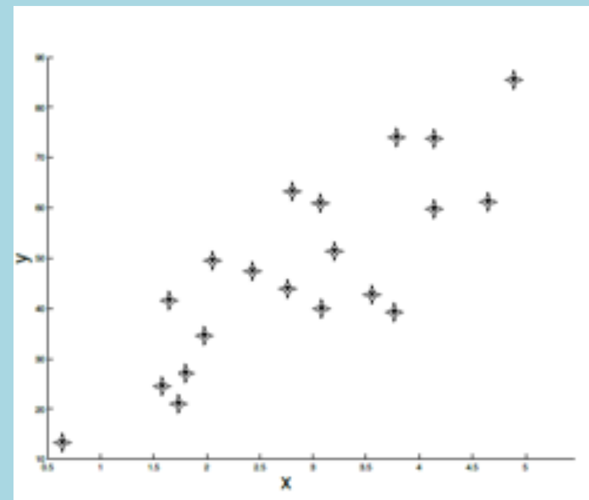
Figure 1: An observed data set and its associated regression line.



(a) Old and new regression lines. (b) Old and new regression lines. (c) Old and new regression lines.

Figure 2: New regression lines for altered data sets S^{new}.

Dataset



(d) Duplicating the original data set.

Sample Questions

3.1 Linear regression

Consider the dataset S plotted in Fig. 1 along with its associated regression line. For each of the altered data sets S^{new} plotted in Fig. 3, indicate which regression line (relative to the original one) in Fig. 2 corresponds to the regression line for the new data set. Write your answers in the table below.

Dataset	(a)	(b)	(c)	(d)	(e)
Regression line					

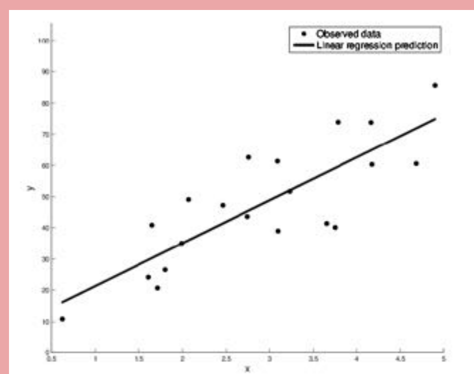
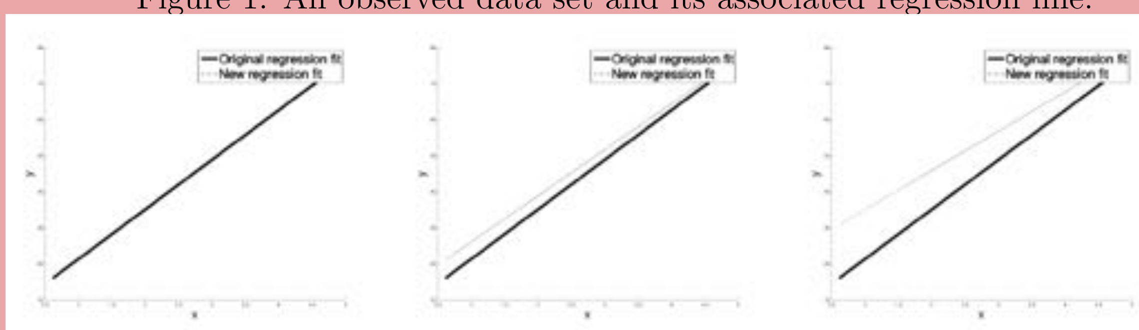


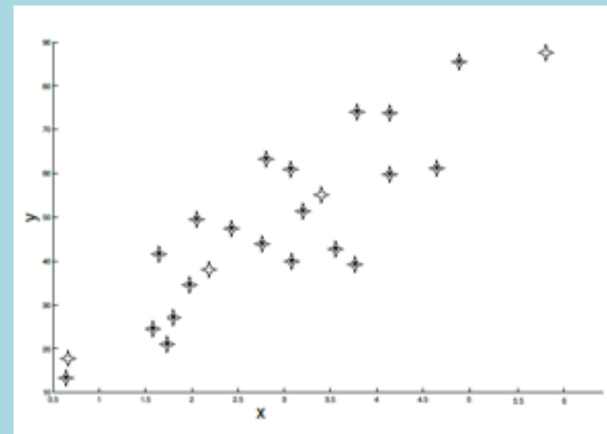
Figure 1: An observed data set and its associated regression line.



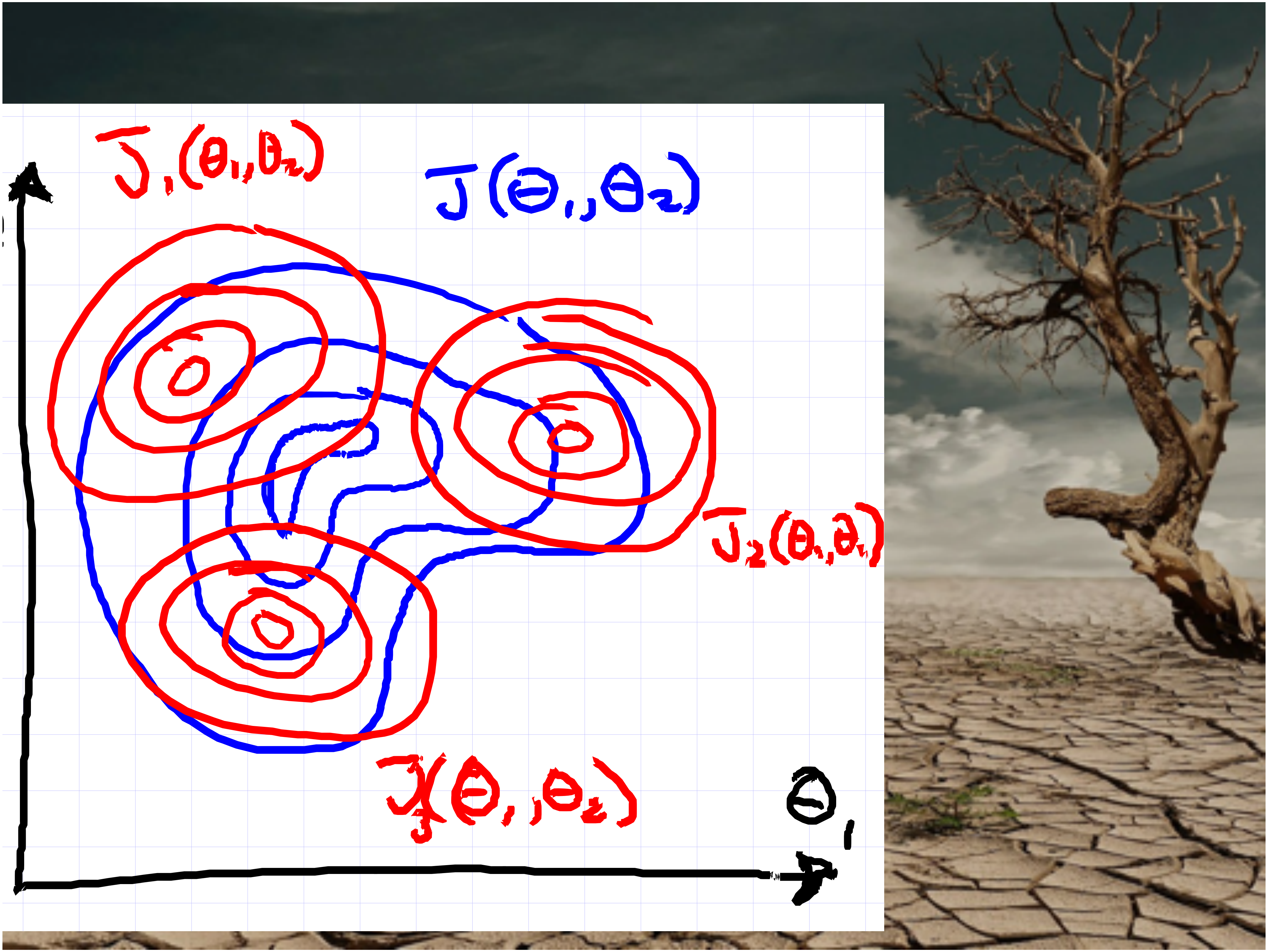
(a) Old and new regression lines. (b) Old and new regression lines. (c) Old and new regression lines.

Figure 2: New regression lines for altered data sets S^{new}.

Dataset



(e) Duplicating the original data set and adding four points that lie on the trajectory of the original regression line.



$J_1(\theta_1,\theta_2)$
$J(\theta_1,\theta_2)$
$J_2(\theta_1,\theta_1)$
$J_3(\theta_1,\theta_2)$
θ_1

Robotic Farming

$y^{(i)} = c^*(x^{(i)})$ $y^{(i)} \sim p^*(\cdot \mid x^{(i)})$

	Deterministic	**Probabilistic**
Classification (binary output)	Is this a picture of a wheat kernel?	Is this plant drought resistant?
Regression (continuous output)	How many wheat kernels are in this picture?	What will the yield of this plant be?

Multinomial Logistic Regression

Sample Questions

3.2 Logistic regression

Given a training set $\{(x_i, y_i), i = 1, \ldots, n\}$ where $x_i \in \mathbb{R}^d$ is a feature vector and $y_i \in \{0, 1\}$ is a binary label, we want to find the parameters $\hat{w}$ that maximize the likelihood for the training set, assuming a parametric model of the form

$$p(y = 1|x; w) = \frac{1}{1 + \exp(-w^T x)}.$$

The conditional log likelihood of the training set is

$$\ell(w) = \sum_{i=1}^{n} y_i \log p(y_i, |x_i; w) + (1 - y_i) \log(1 - p(y_i, |x_i; w)),$$

and the gradient is

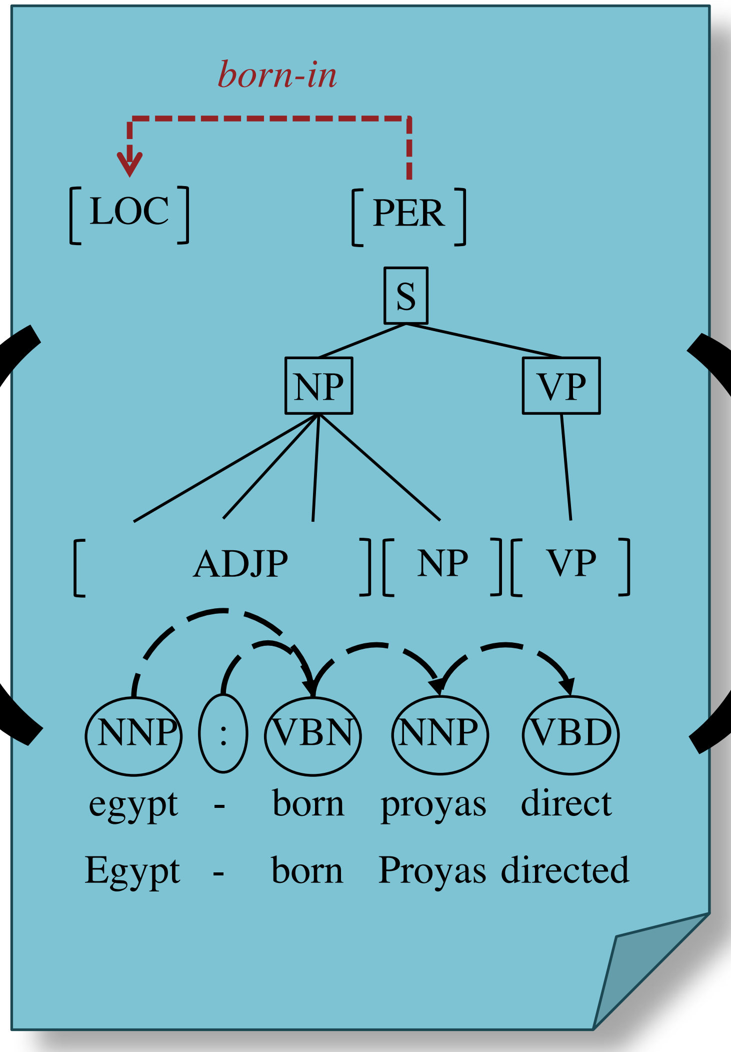
$$\nabla \ell(w) = \sum_{i=1}^{n} (y_i - p(y_i|x_i; w)) x_i.$$

(b) [5 pts.] What is the form of the classifier output by logistic regression?

(c) [2 pts.] **Extra Credit:** Consider the case with binary features, i.e, $x \in \{0, 1\}^d \subset \mathbb{R}^d$, where feature x_1 is rare and happens to appear in the training set with only label 1. What is $\hat{w}_1$? Is the gradient ever zero for any finite w? Why is it important to include a regularization term to control the norm of $\hat{w}$?

Handcrafted Features

$p(y|x) \propto \exp(\Theta_y \bullet f\Big($



$\Big))$

Example: Linear Regression

Goal: Learn $y = \mathbf{w}^T f(\mathbf{x}) + b$ where f(.) is a polynomial basis function

true "unknown" target function is linear with negative slope and gaussian noise

Regularization

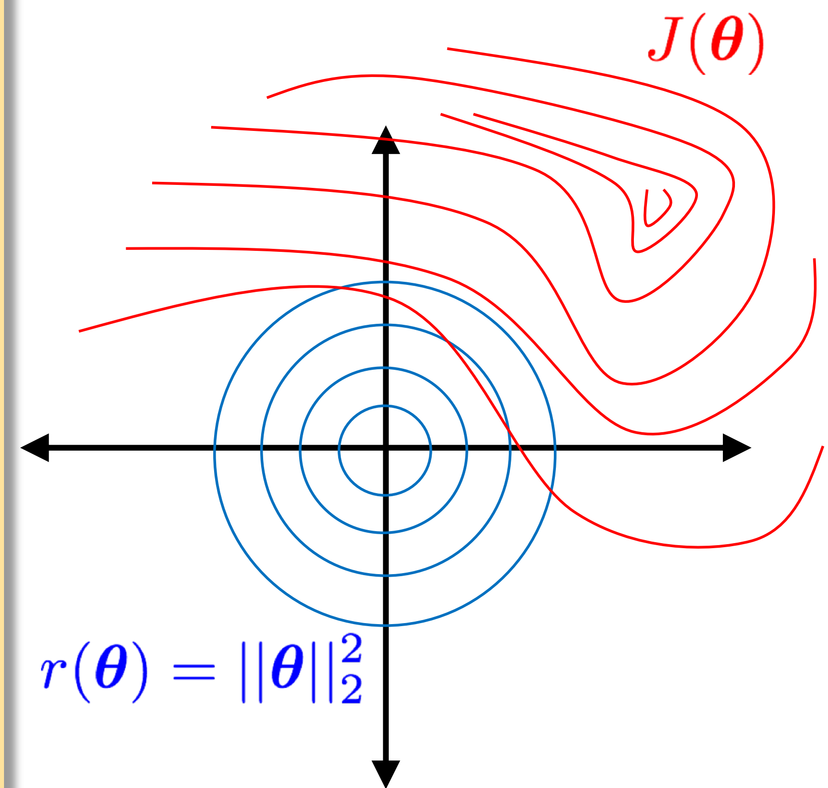
Question:

Suppose we are minimizing J'(θ) where

$$\boxed{J'(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) + \lambda r(\boldsymbol{\theta})}$$

As λ increases, the minimum of J'(θ) will…

A. … move towards the midpoint between J'(θ) and r(θ)
B. … move towards the minimum of J(θ)
C. … move towards the minimum of r(θ)
D. … move towards a theta vector of positive infinities
E. … move towards a theta vector of negative infinities
F. … stay the same



Samples Questions

2.1 Train and test errors

In this problem, we will see how you can debug a classifier by looking at its train and test errors. Consider a classifier trained till convergence on some training data $\mathcal{D}^{\text{train}}$, and tested on a separate test set $\mathcal{D}^{\text{test}}$. You look at the test error, and find that it is very high. You then compute the training error and find that it is close to 0.

1. **[4 pts]** Which of the following is expected to help? Select all that apply.

   (a) Increase the training data size.

   (b) Decrease the training data size.

   (c) Increase model complexity (For example, if your classifier is an SVM, use a more complex kernel. Or if it is a decision tree, increase the depth).

   (d) Decrease model complexity.

   (e) Train on a combination of $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{test}}$ and test on $\mathcal{D}^{\text{test}}$

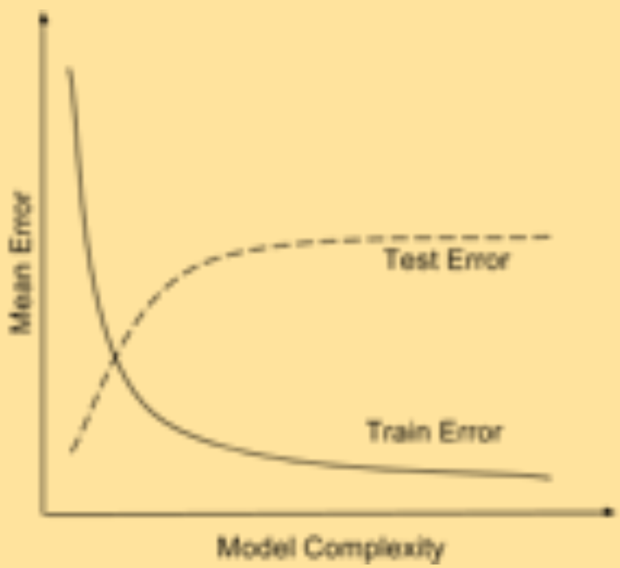
   (f) Conclude that Machine Learning does not work.

Samples Questions

2.1 Train and test errors

In this problem, we will see how you can debug a classifier by looking at its train and test errors. Consider a classifier trained till convergence on some training data $\mathcal{D}^{\text{train}}$, and tested on a separate test set $\mathcal{D}^{\text{test}}$. You look at the test error, and find that it is very high. You then compute the training error and find that it is close to 0.

4. **[1 pts]** Say you plot the train and test errors as a function of the model complexity. Which of the following two plots is your plot expected to look like?



(a)

(b)

Sample Questions

4.1 True or False

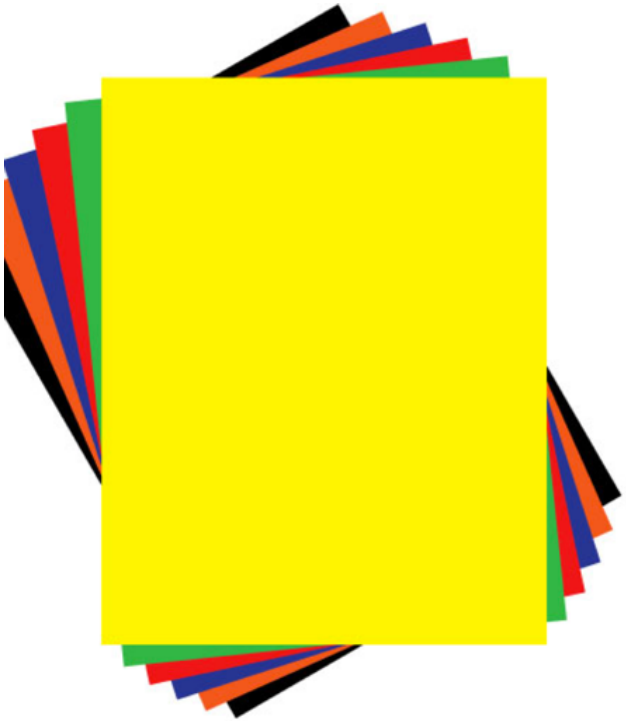
Answer each of the following questions with **T** or **F** and **provide a one line justification**.

(a) [2 pts.] Consider two datasets $D^{(1)}$ and $D^{(2)}$ where $D^{(1)} = \{(x_1^{(1)}, y_1^{(1)}), ..., (x_n^{(1)}, y_n^{(1)})\}$ and $D^{(2)} = \{(x_1^{(2)}, y_1^{(2)}), ..., (x_m^{(2)}, y_m^{(2)})\}$ such that $x_i^{(1)} \in \mathbb{R}^{d_1}$, $x_i^{(2)} \in \mathbb{R}^{d_2}$. Suppose $d_1 > d_2$ and $n > m$. Then the maximum number of mistakes a perceptron algorithm will make is higher on dataset $D^{(1)}$ than on dataset $D^{(2)}$.

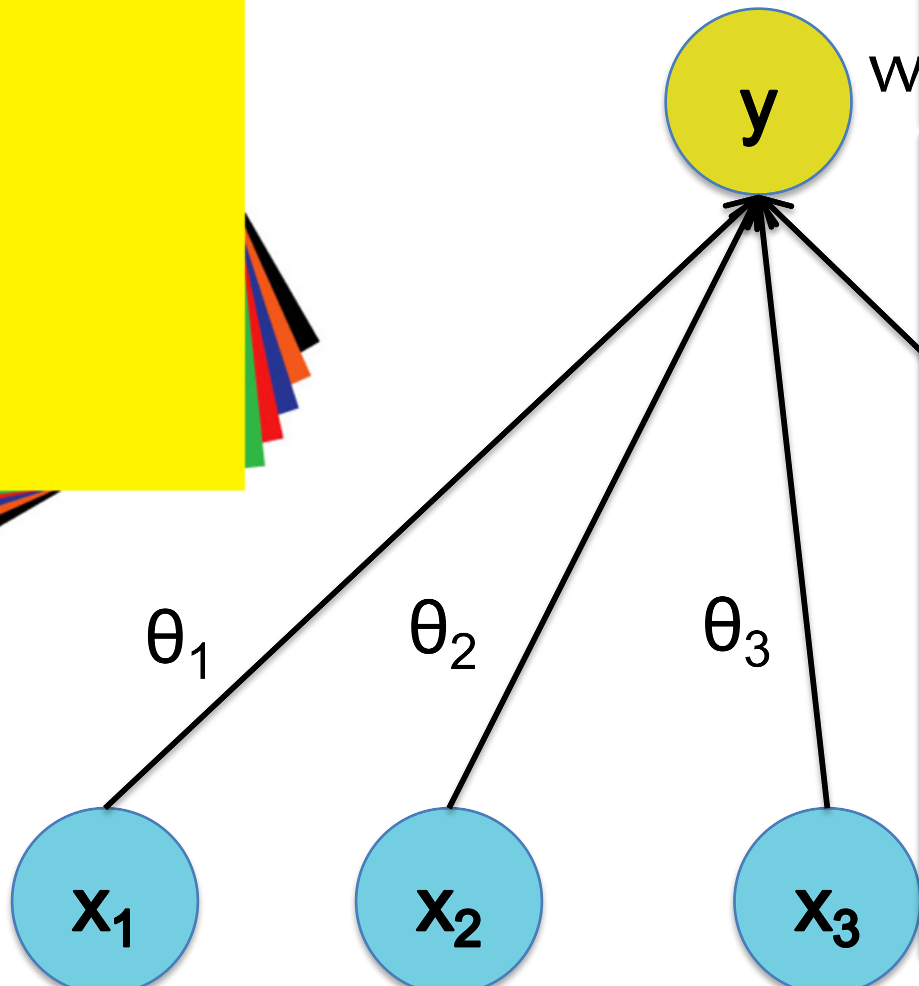
Decision Functions

Logistic Regression

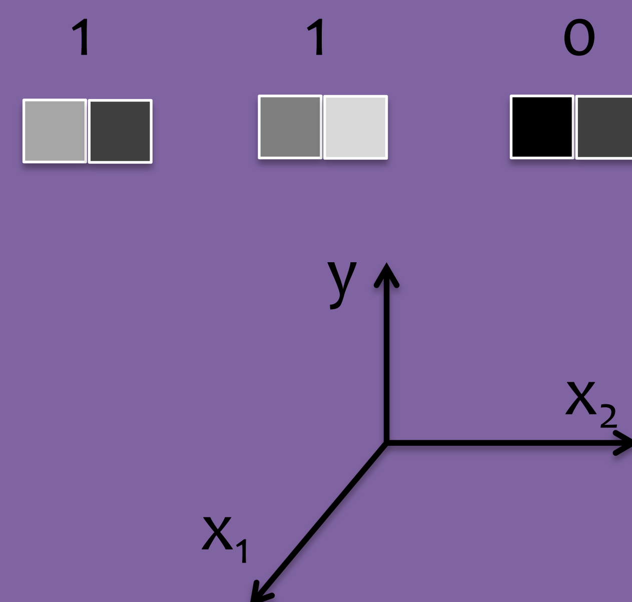
$$y = h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^T \boldsymbol{x})$$



Input



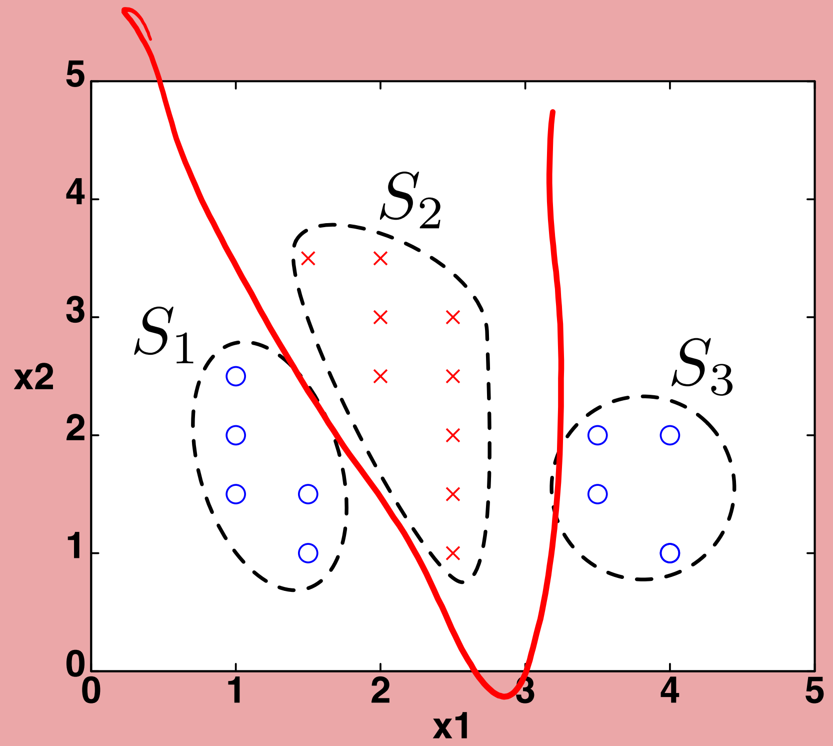
In-Class Example



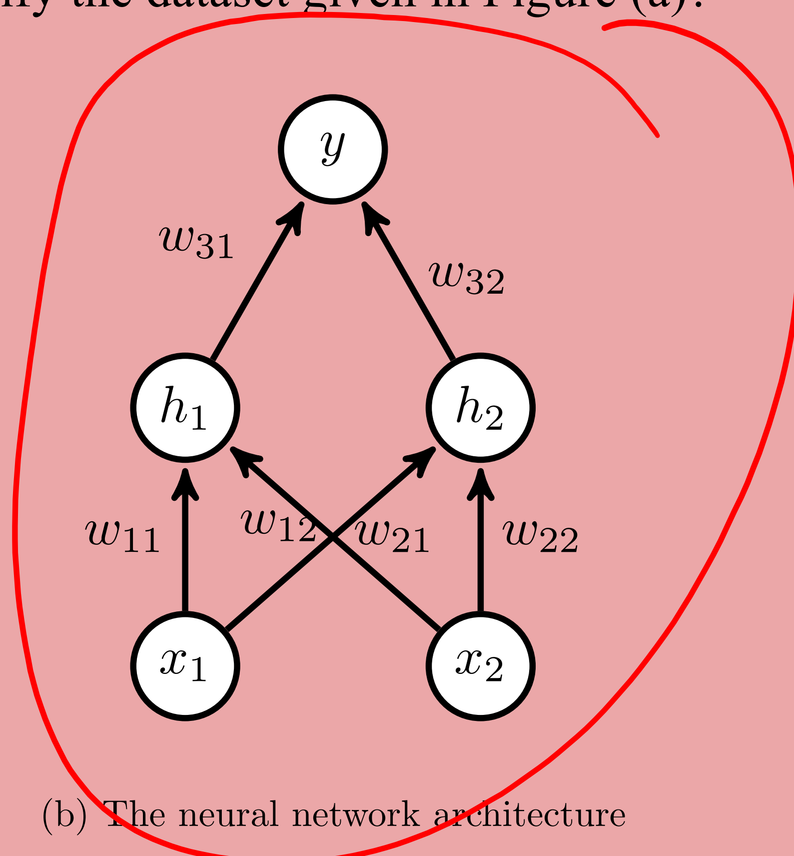
Sample Questions

Neural Networks

Can the neural network in Figure (b) correctly classify the dataset given in Figure (a)?



(a) The dataset with groups S_1, S_2, and S_3.

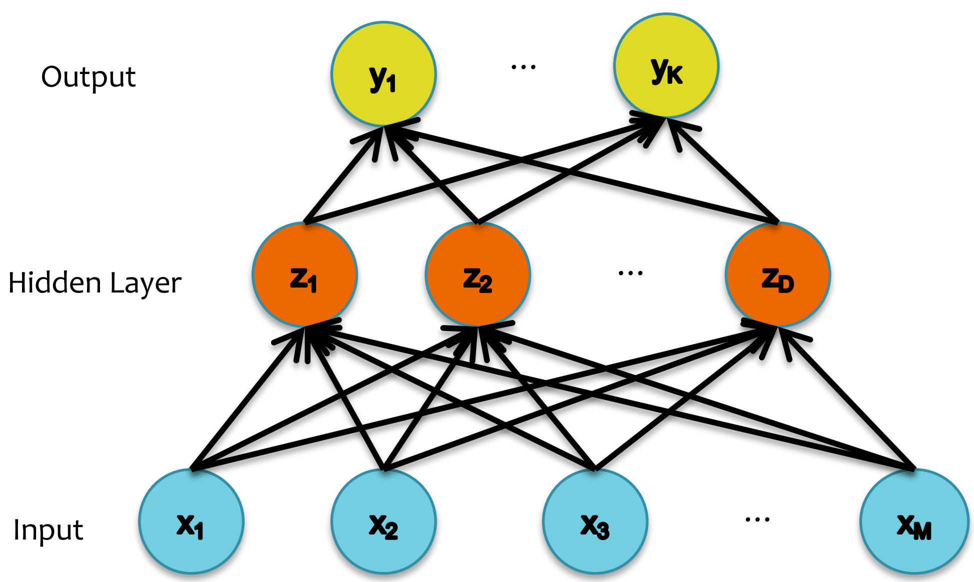


(b) The neural network architecture

Multi-Class Output

Softmax:

$$y_k = \frac{\exp(b_k)}{\sum_{l=1}^{K} \exp(b_l)}$$



(F) Loss
$J = \sum_{k=1}^{K} y_k^* \log(y_k)$

(E) Output (softmax)
$y_k = \frac{\exp(b_k)}{\sum_{l=1}^{K} \exp(b_l)}$

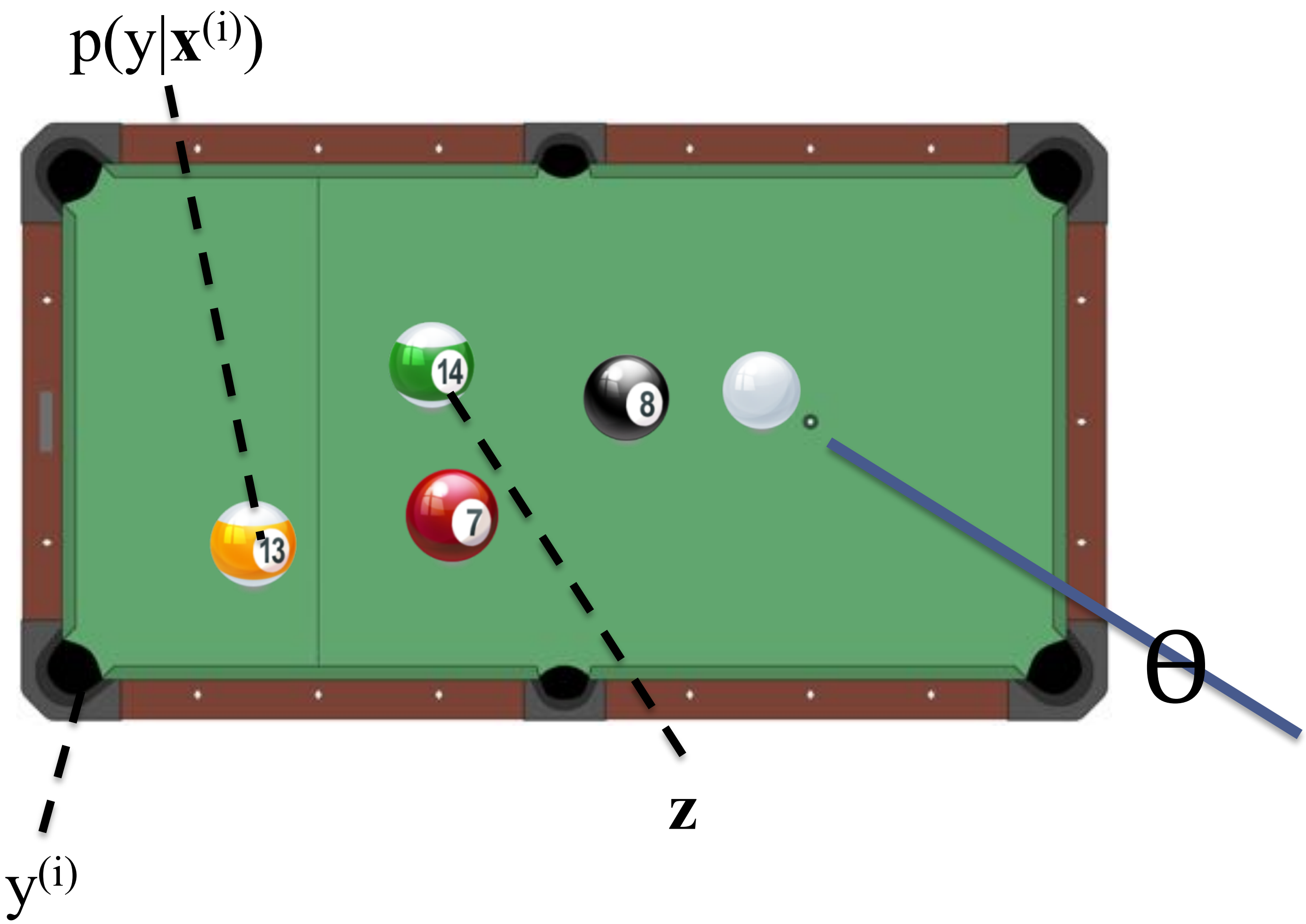
(D) Output (linear)
$b_k = \sum_{j=0}^{D} \beta_{kj} z_j \;\; \forall k$

(C) Hidden (nonlinear)
$z_j = \sigma(a_j), \;\; \forall j$

(B) Hidden (linear)
$a_j = \sum_{i=0}^{M} \alpha_{ji} x_i, \;\; \forall j$

(A) Input
Given $x_i, \;\; \forall i$

Error Back-Propagation

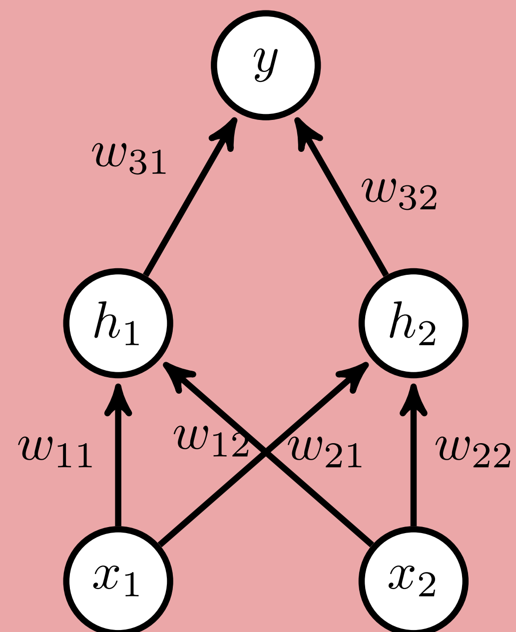


Slide from (Stoyanov & Eisner, 2012)

Sample Questions

Neural Networks

Apply the backpropagation algorithm to obtain the partial derivative of the mean-squared error of y with the true value y^* with respect to the weight w_{22} assuming a sigmoid nonlinear activation function for the hidden layer.



(b) The neural network architecture

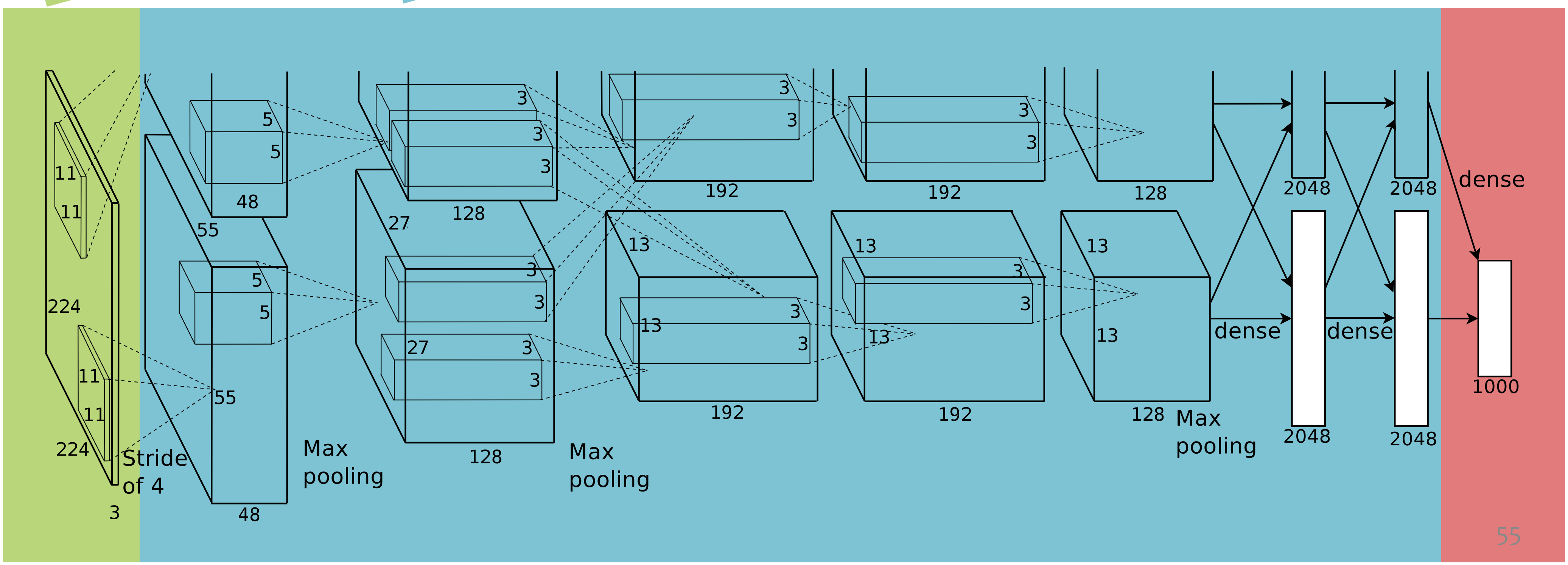
Architecture #2: AlexNet

CNN for Image Classification
(Krizhevsky, Sutskever & Hinton, 2012)
15.3% error on ImageNet LSVRC-2012 contest

Input image (pixels)

- Five convolutional layers (w/max-pooling)
- Three fully connected layers

1000-way softmax



Bidirectional RNN

inputs: $\mathbf{x} = (x_1, x_2, \ldots, x_T), x_i \in \mathcal{R}^I$

hidden units: $\overrightarrow{\mathbf{h}}$ and $\overleftarrow{\mathbf{h}}$

outputs: $\mathbf{y} = (y_1, y_2, \ldots, y_T), y_i \in \mathcal{R}^K$

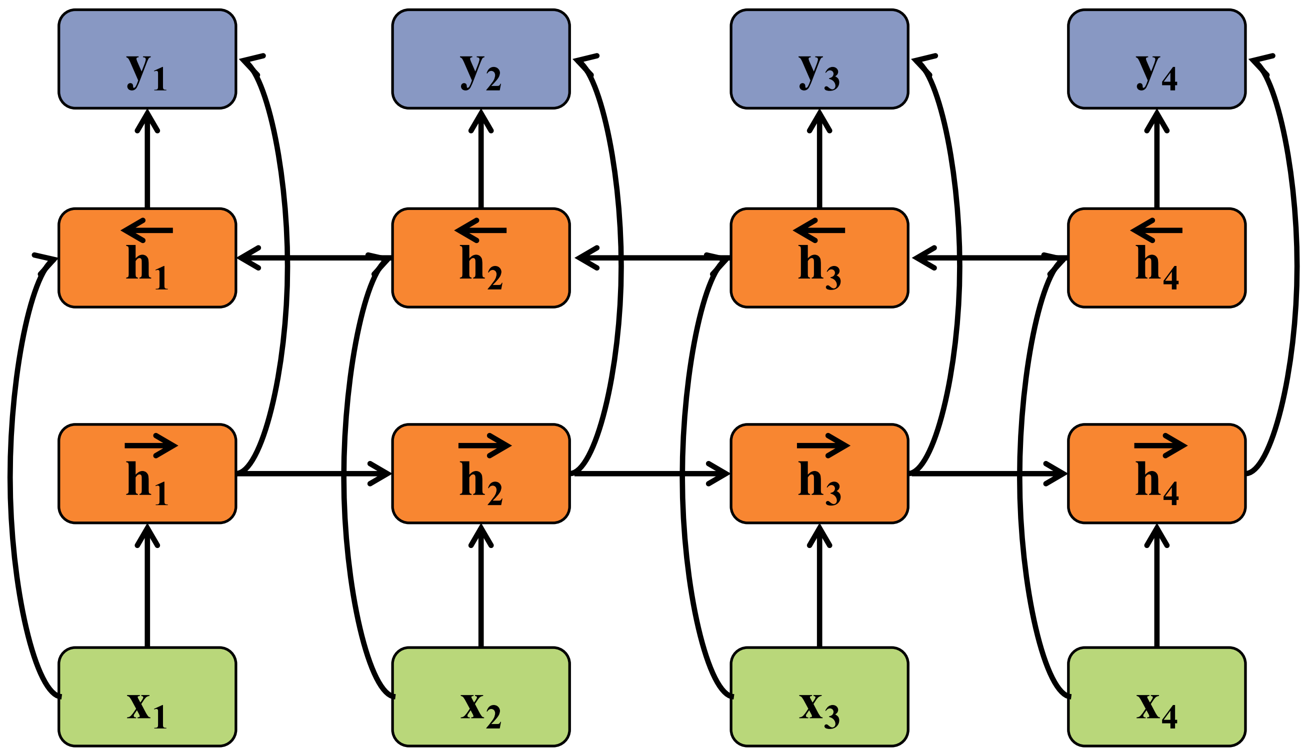
nonlinearity: $\mathcal{H}$

Recursive Definition:

$$\overrightarrow{h}_t = \mathcal{H}\left(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right)$$

$$\overleftarrow{h}_t = \mathcal{H}\left(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}\right)$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y$$



PAC-MAN Learning

For some hypothesis $h \in \mathcal{H}$:

1. True Error

$$R(h)$$

2. Training Error

$$\hat{R}(h)$$

Question 2:
What is the expected number of PAC-MAN levels Matt will complete before a **Game-Over**?

A. 1-10
B. 11-20
C. 21-30

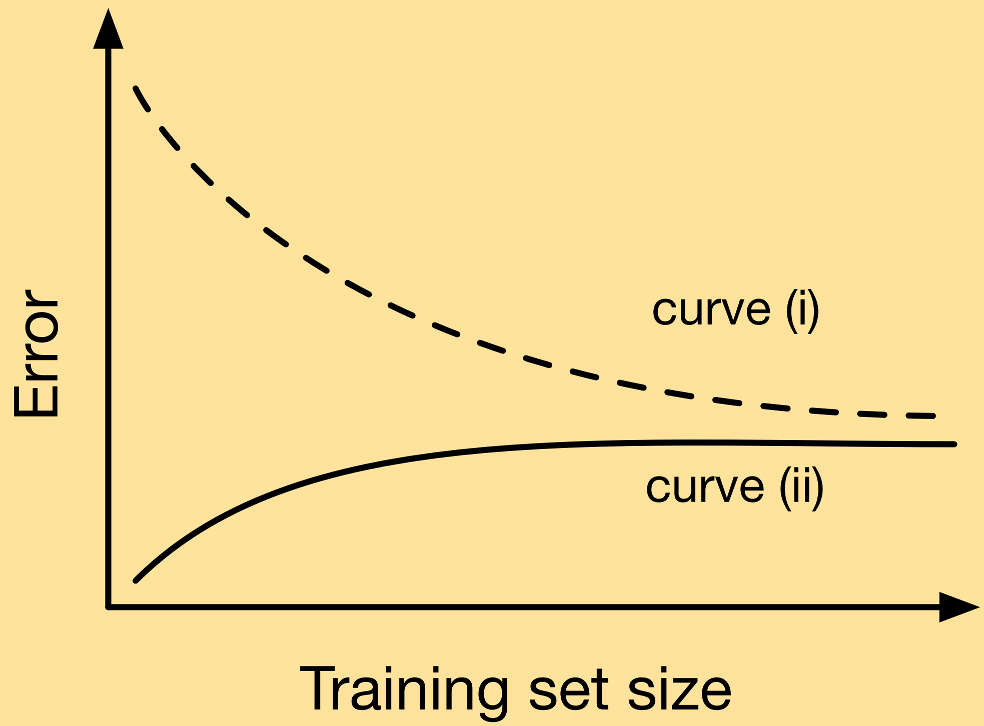
Samples Questions

2.1 True Errors

(b) [4 pts.] **T or F:** Learning theory allows us to determine with 100% certainty the true error of a hypothesis to within any $\epsilon > 0$ error.

Samples Questions

2.2 Training Sample Size



(a) [8 pts.] Which curve represents the training error? **Please provide 1–2 sentences of justification**.

(b) [4 pt.] In one word, what does the gap between the two curves represent?

Sample Questions

5 Learning Theory [20 pts.]

(a) [3 pts.] **T or F**: It is possible to label 4 points in $\mathbb{R}^2$ in all possible 2^4 ways via linear separators in $\mathbb{R}^2$.

(d) [3 pts.] **T or F**: The VC dimension of a concept class with infinite size is also infinite.

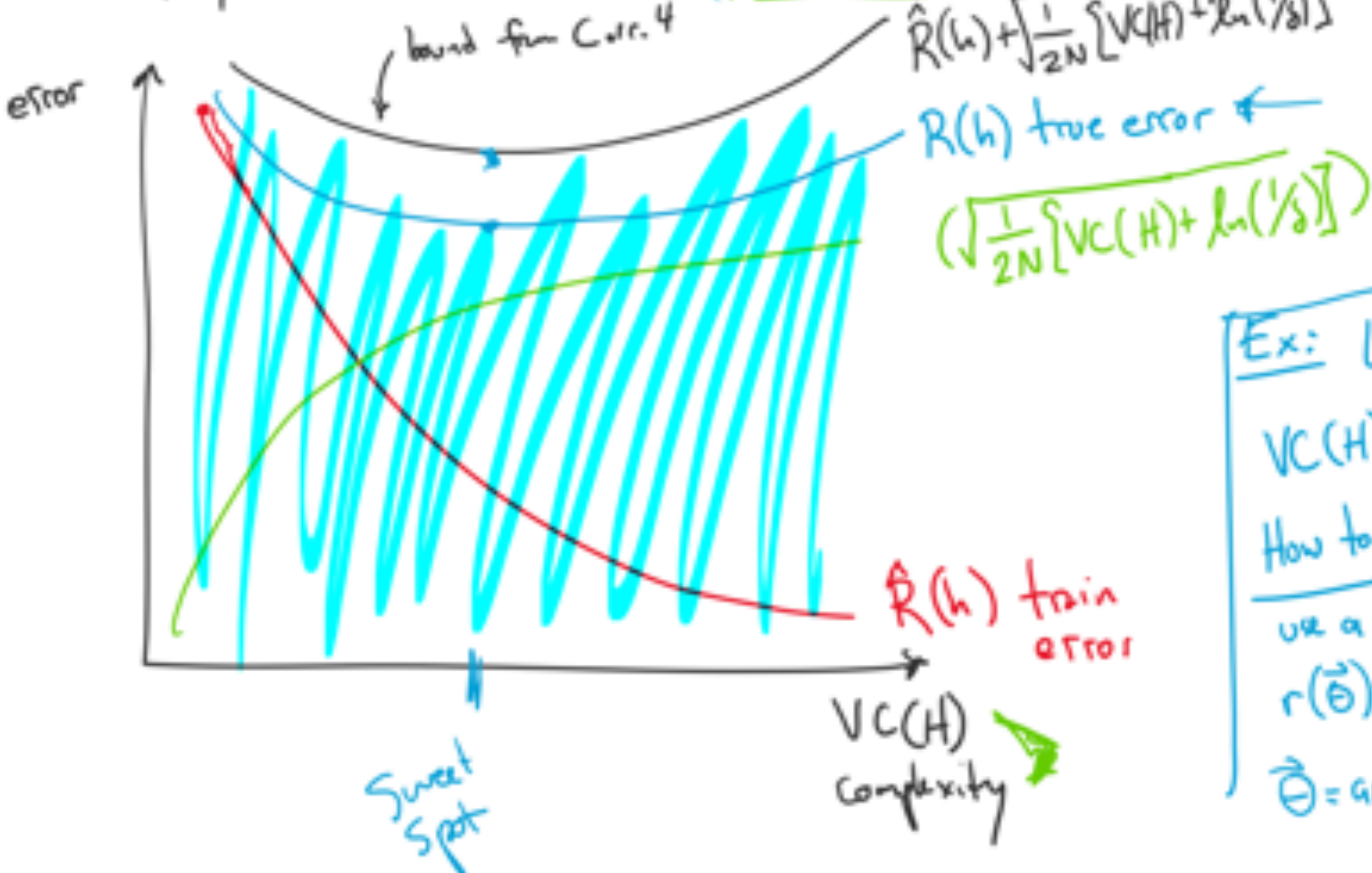
(f) [3 pts.] **T or F**: Given a realizable concept class and a set of training instances, a consistent learner will output a concept that achieves 0 error on the training instances.

PAC Learning & Regularization

Model Selection

Q. Is Corr. 4 useful? A: Yes!

Key Idea: tradeoff between low train error and keeping H simple (low VCDim)



Ex: Lin. Sep. in $\mathbb{R}^M$

$VC(H) = M+1$

How to tradeoff?

use a regularizer!

$r(\vec{\theta}) = \sum_{m=1}^{M} |\theta_m|$

$\vec{\theta} = \underset{\theta}{\operatorname{argmin}}\ J(\vec{\theta}) + r(\vec{\theta})$

MLE vs. MAP

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

Principle of Maximum Likelihood Estimation:
Choose the parameters that maximize the likelihood of the data.

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)

Principle of Maximum *a posteriori* (MAP) Estimation:
Choose the parameters that maximize the posterior of the parameters given the data.

$$\boldsymbol{\theta}^{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})\overbrace{p(\boldsymbol{\theta})}^{\text{Prior}}$$

Maximum *a posteriori* (MAP) estimate

Sample Questions

1.2 Maximum Likelihood Estimation (MLE)

Assume we have a random sample that is Bernoulli distributed $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$. We are going to derive the MLE for θ. Recall that a Bernoulli random variable X takes values in $\{0, 1\}$ and has probability mass function given by

$$P(X; \theta) = \theta^X (1 - \theta)^{1-X}.$$

(a) [2 pts.] Derive the likelihood, $L(\theta; X_1, \ldots, X_n)$.

(c) **Extra Credit:** [2 pts.] Derive the following formula for the MLE: $\hat{\theta} = \dfrac{1}{n} \left(\sum_{i=1}^n X_i\right)$.

Sample Questions

1.3 MAP vs MLE

Answer each question with **T** or **F** and **provide a one sentence explanation of your answer**:

(a) [2 pts.] **T or F:** In the limit, as n (the number of samples) increases, the MAP and MLE estimates become the same.

Fake News Detector

Today's Goal: To define a generative model of emails of two different classes (e.g. real vs. fake news)

The Economist

La paralización

Spain may be heading for its third election in a year

All latest updates

Stubborn Socialists are blocking Mariano Rajoy from forming a centre-right government

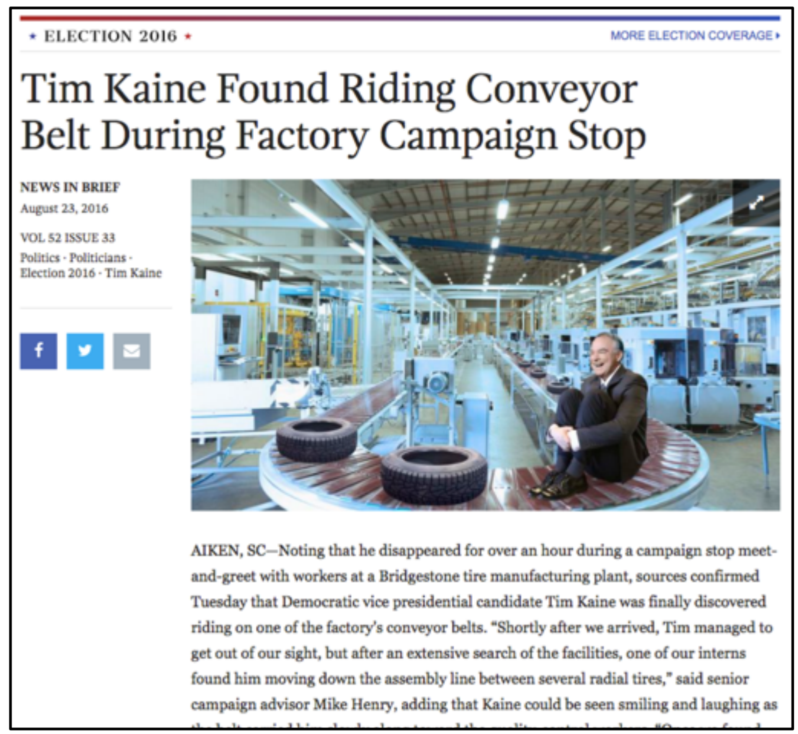
Sep 5th 2016 | MADRID | Europe

Timekeeper · Like 80 · Tweet

EPA

BACK in June, after Spain's second indecisive election in six months, the general expectation was that Mariano Rajoy, the prime minister, would swiftly form a new government. Although his conservative People's Party (PP) did not win back the absolute majority it had lost in December, it remained easily the largest party, with 137 of the 350

The Onion



ELECTION 2016

MORE ELECTION COVERAGE

Tim Kaine Found Riding Conveyor Belt During Factory Campaign Stop

NEWS IN BRIEF

August 23, 2016

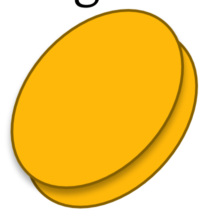
VOL 52 ISSUE 33

Politics · Politicians · Election 2016 · Tim Kaine

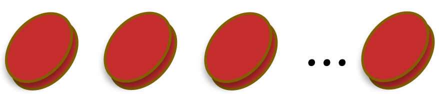
AIKEN, SC—Noting that he disappeared for over an hour during a campaign stop meet-and-greet with workers at a Bridgestone tire manufacturing plant, sources confirmed Tuesday that Democratic vice presidential candidate Tim Kaine was finally discovered riding on one of the factory's conveyor belts. "Shortly after we arrived, Tim managed to get out of our sight, but after an extensive search of the facilities, one of our interns found him moving down the assembly line between several radial tires," said senior campaign advisor Mike Henry, adding that Kaine could be seen smiling and laughing as

Model 1: Bernoulli Naïve Bayes

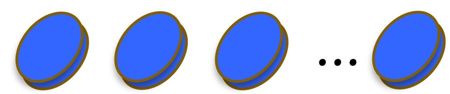
Flip weighted coin



If HEADS, flip each red coin



If TAILS, flip each blue coin



y	x_1	x_2	x_3	…	x_M
0	1	0	1	…	1
1	0	1	0	…	1
1	1	1	1	…	1
0	0	0	1	…	1
0	1	0	1	…	0
1	1	0	1	…	0

We can **generate** data in this fashion. Though in practice we never would since our data is **given.**

Instead, this provides an explanation of **how** the data was generated (albeit a terrible one).

Each red coin corresponds to an x_m

Sample Questions

1.1 Naive Bayes

You are given a data set of 10,000 students with their sex, height, and hair color. You are trying to build a classifier to predict the sex of a student, so you randomly split the data into a training set and a testing set. Here are the specifications of the data set:

- sex $\in$ {male,female}
- height $\in$ [0,300] centimeters
- hair $\in$ {brown, black, blond, red, green}
- 3240 men in the data set
- 6760 women in the data set

Under the assumptions necessary for Naive Bayes (not the distributional assumptions you might naturally or intuitively make about the dataset) answer each question with **T** or **F** and **provide a one sentence explanation of your answer**:

(a) [2 pts.] **T or F:** As height is a continuous valued variable, Naive Bayes is not appropriate since it cannot handle continuous valued variables.

(c) [2 pts.] **T or F:** $P(\texttt{height}|\texttt{sex}, \texttt{hair}) = P(\texttt{height}|\texttt{sex})$.

Material Covered **After** Midterm Exam 2

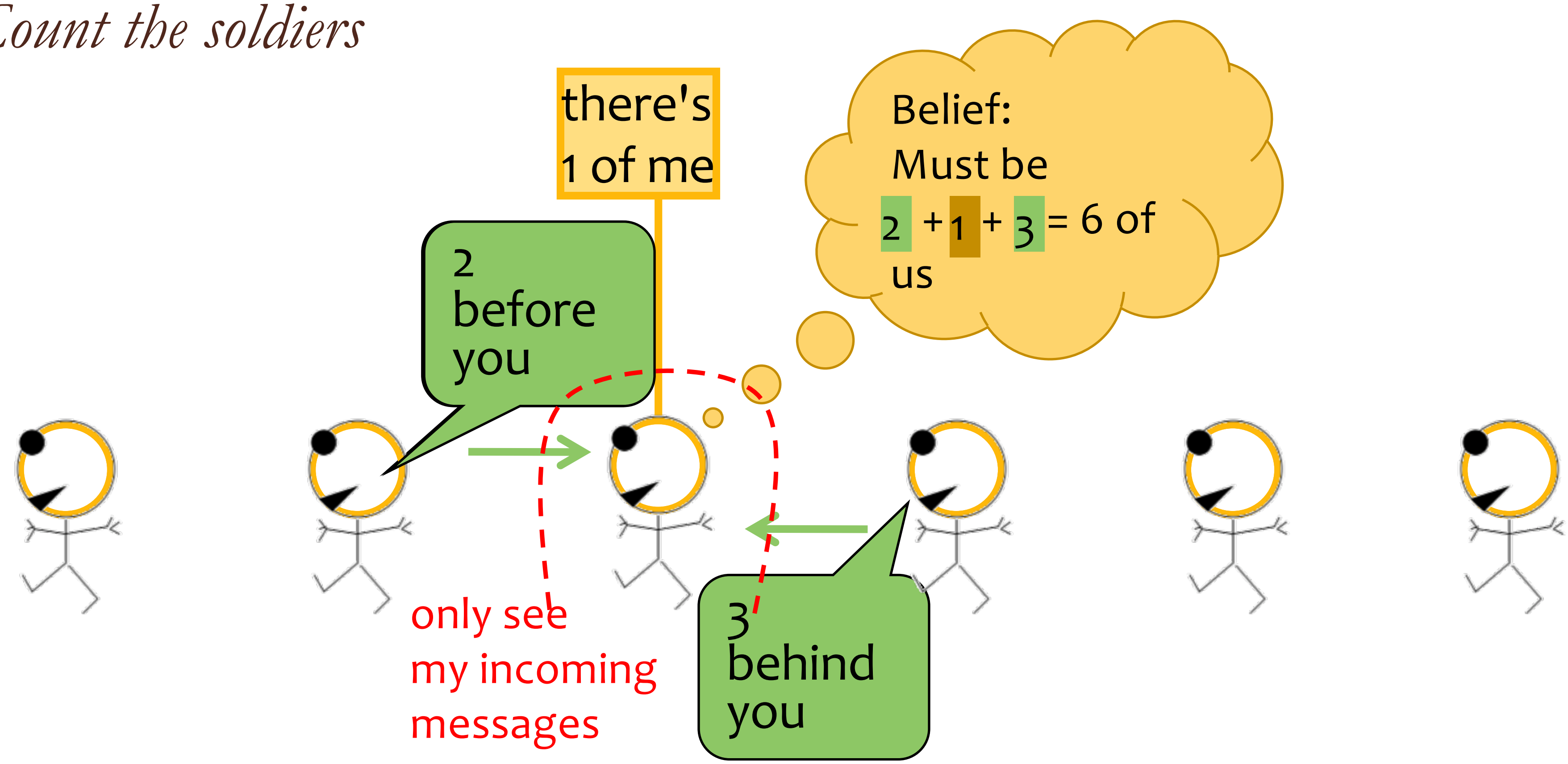
SAMPLE QUESTIONS

Totoro’s Tunnel

SQUIRREL
HILL SOUTH
Covode St
Nicholson St
Phillips Ave
Crombie St
Melvin St
Taylor Allderdice
High School
Tilbury Ave
Beechwood Blvd
English Ln
Mt Royal Rd
376
Beechwood Blvd
Alderson St
Forward Ave
Forward Ave
Mirror St
Flemington St
Inez Way
Burchfield Ave
Inez Way
376
Burchfield Ave
Ira Way
Anita Ave
376
Kennebec St
Monitor St
Monitor St
Summerset Dr
Beechwood Blvd
Saline St
Shady Ave
Ebdy St
Fernwald Rd
Windsor St
Murray Ave
Caton St
Nine Mile Run
Lilac St
Rosemoor St
Rosemont Ln
Melbourne St
Landview Rd
Map data ©2018 Google
US FOODS

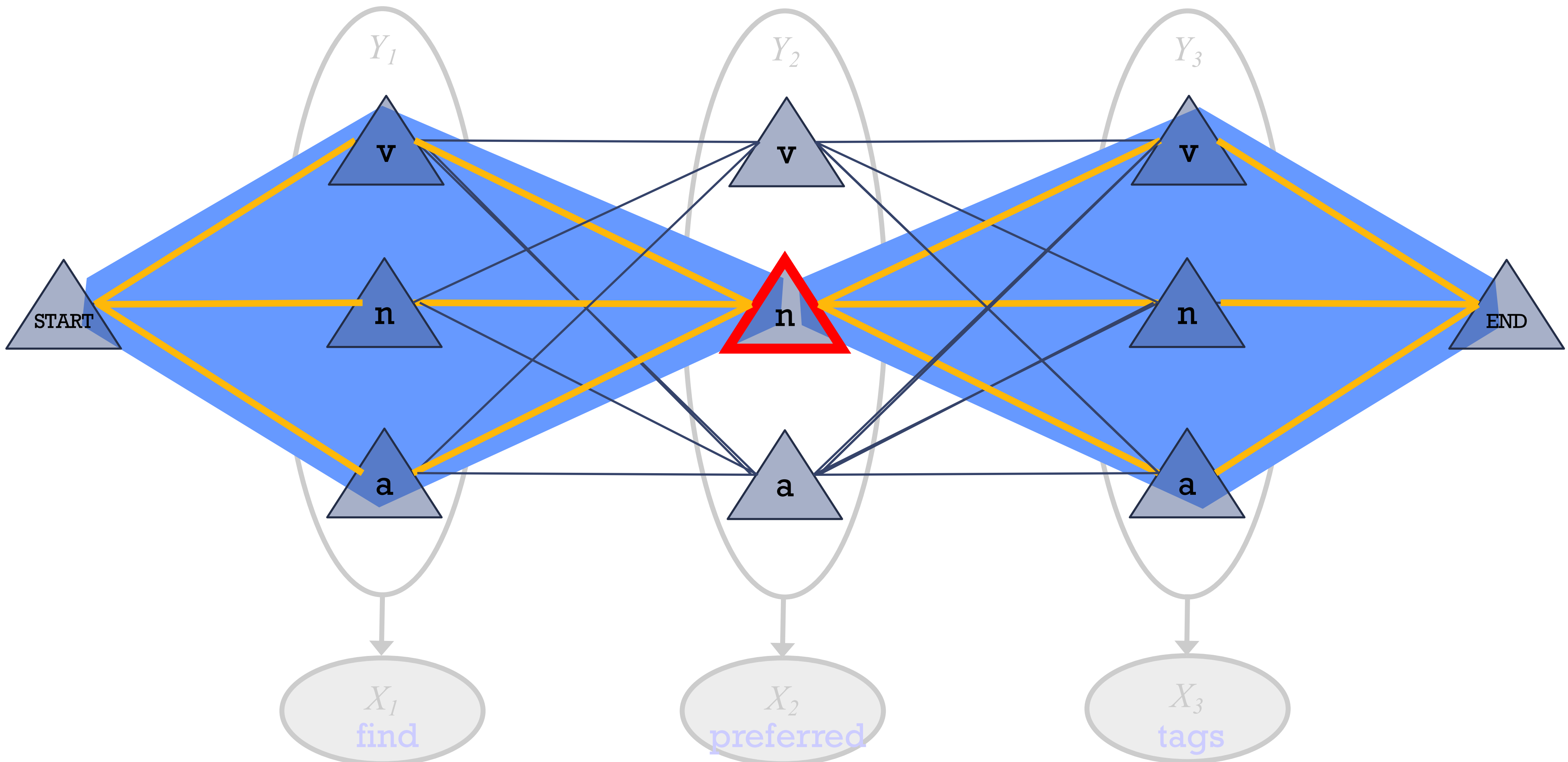
Great Ideas in ML: Message Passing

Count the soldiers



adapted from MacKay (2003) textbook

Forward-Backward Algorithm: Finds Marginals



$\alpha_2(\mathbf{n})$ = total weight of these path *prefixes* (a + b + c)

$\beta_2(\mathbf{n})$ = total weight of these path *suffixes* (x + y + z)

Product gives ax+ay+az+bx+by+bz+cx+cy+cz = total weight of paths

Sample Questions

4 Hidden Markov Models

1. Given the POS tagging data shown, what are the parameter values learned by an HMM?

Verb	**Noun**	**Verb**
see	spot	run

Verb	**Noun**	**Verb**
run	spot	run

Adj.	**Adj.**	**Noun**
funny	funny	spot

$A =$

	Verb	Noun	Adj
see	1/4		
spot	0/4		
run	3/4		
funny	0/4		

$p(x_t \mid y_t)$

$C =$

Verb	Noun	Adj
2/3	0/3	1/3

$B =$

	Verb	Noun	Adj
V			
N			
A			

Sample Questions

4 Hidden Markov Models

1. Given the POS tagging data shown, what are the parameter values learned by an HMM?

2. Suppose you a learning an HMM POS Tagger, how many POS tag sequences of length 23 are there? given 45 pos tags

3. How does an HMM efficiently search for the most probable tag sequence given a 23 word sentence?

Viterbi Algo.

Verb	Noun	Verb
see	spot	run

Verb	Noun	Verb
run	spot	run

Adj.	Adj.	Noun
funny	funny	spot

Example: Ryan Reynolds' Voicemail

From https://www.adweek.com/brand-marketing/ryan-reynolds-left-voicemails-for-all-mint-mobile-subscribers/

Example: Tornado Alarms

Hacking Attack Woke Up Dallas With Emergency Sirens, Officials Say

By ELI ROSENBERG and MAYA SALAM APRIL 8, 2017

Warning sirens in Dallas, meant to alert the public to emergencies like severe weather, started sounding around 11:40 p.m. Friday, and were not shut off until 1:20 a.m. Rex C. Curry for The New York Times

1. Imagine that you work at the 911 call center in Dallas
2. You receive six calls informing you that the Emergency Weather Sirens are going off
3. What do you conclude?

Figure from https://www.nytimes.com/2017/04/08/us/dallas-emergency-sirens-hacking.html

Sample Questions

(a) [2 pts.] Write the expression for the joint distribution.

$p(S,R,E,A) = \underline{p(A|E)}\,\underline{p(E|S,R)}\,\underline{p(S)}\,\underline{p(R)}$

5 Graphical Models [16 pts.]

We use the following Bayesian network to model the relationship between studying (S), being well-rested (R), doing well on the exam (E), and getting an A grade (A). All nodes are binary, i.e., $R, S, E, A \in \{0, 1\}$.

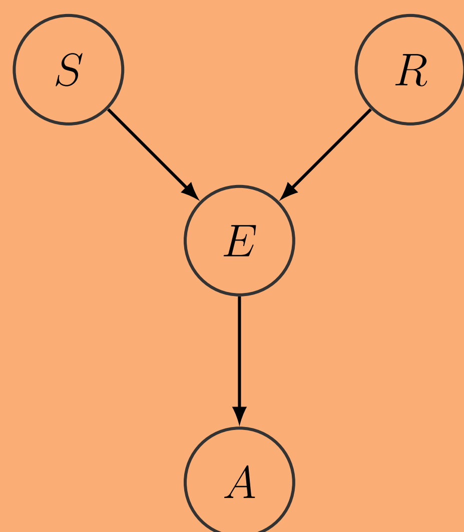


Figure 5: Directed graphical model for problem 5.

poll.mlcourse.97

Sample Questions

Q1:

(b) [2 pts.] How many parameters, i.e., entries in the CPT tables, are necessary to describe the joint distribution?

16

5 Graphical Models [16 pts.]

We use the following Bayesian network to model the relationship between studying (S), being well-rested (R), doing well on the exam (E), and getting an A grade (A). All nodes are binary, i.e., $R, S, E, A \in \{0, 1\}$.

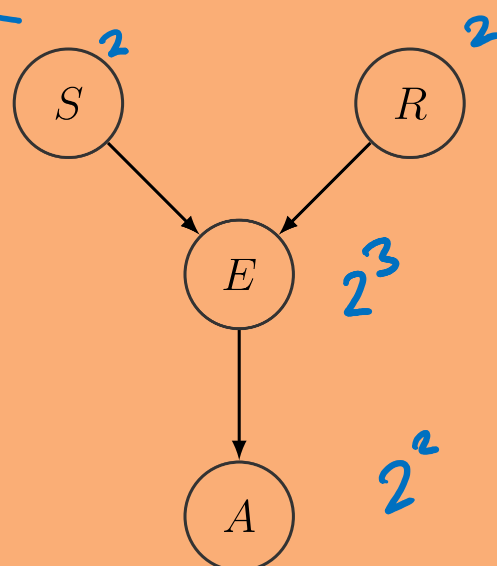


Figure 5: Directed graphical model for problem 5.

Sample Questions

(d) [2 pts.] Is S marginally independent of R? Is S conditionally independent of R given E? Answer yes or no to each questions and provide a brief explanation why.

5 Graphical Models [16 pts.]

We use the following Bayesian network to model the relationship between studying (S), being well-rested (R), doing well on the exam (E), and getting an A grade (A). All nodes are binary, i.e., $R, S, E, A \in \{0, 1\}$.

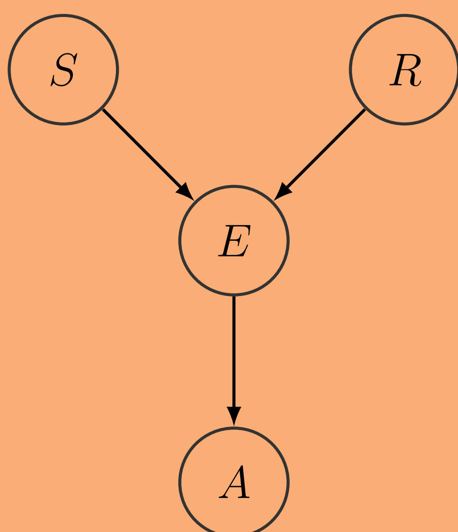


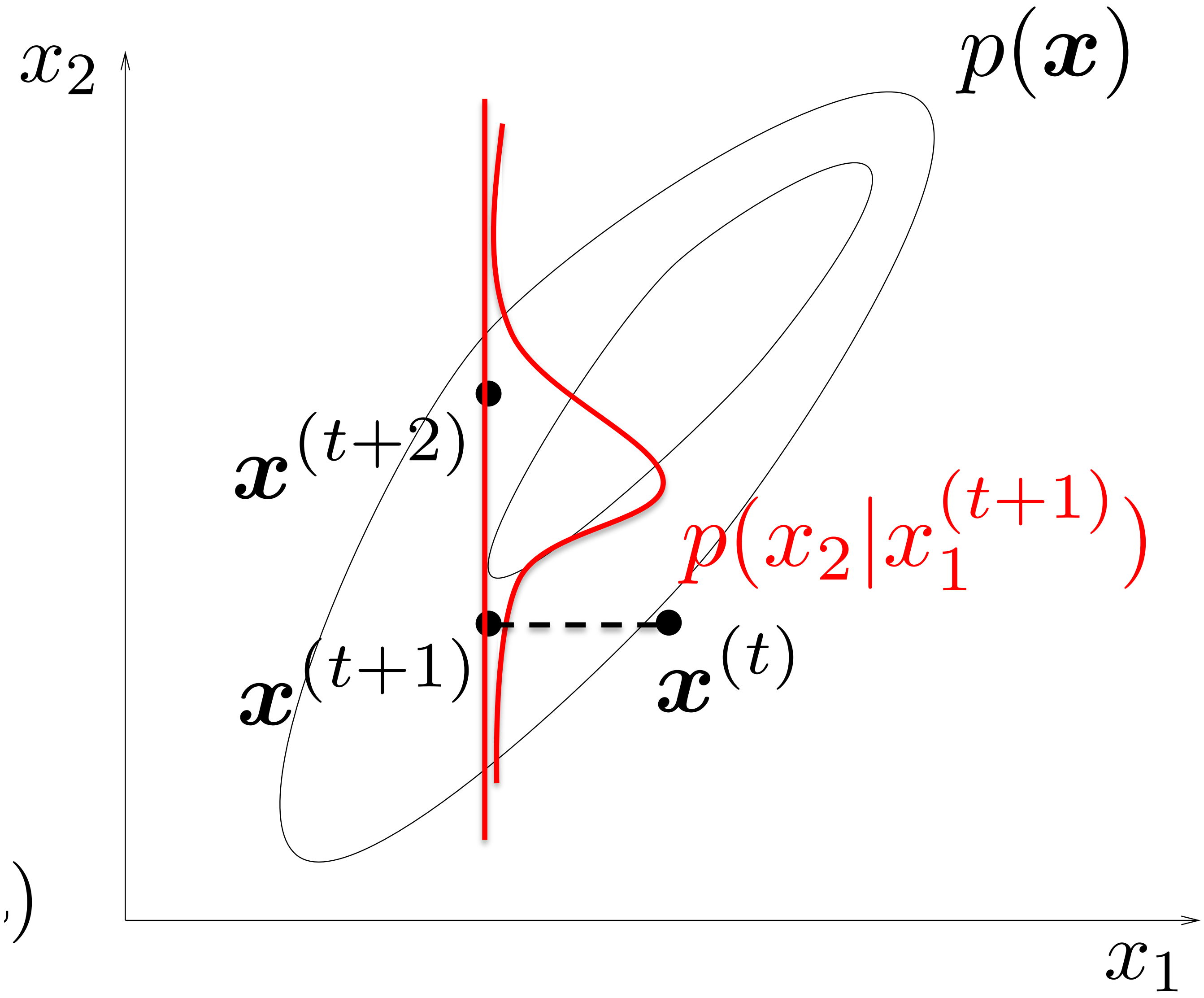
Figure 5: Directed graphical model for problem 5.

Sample Questions

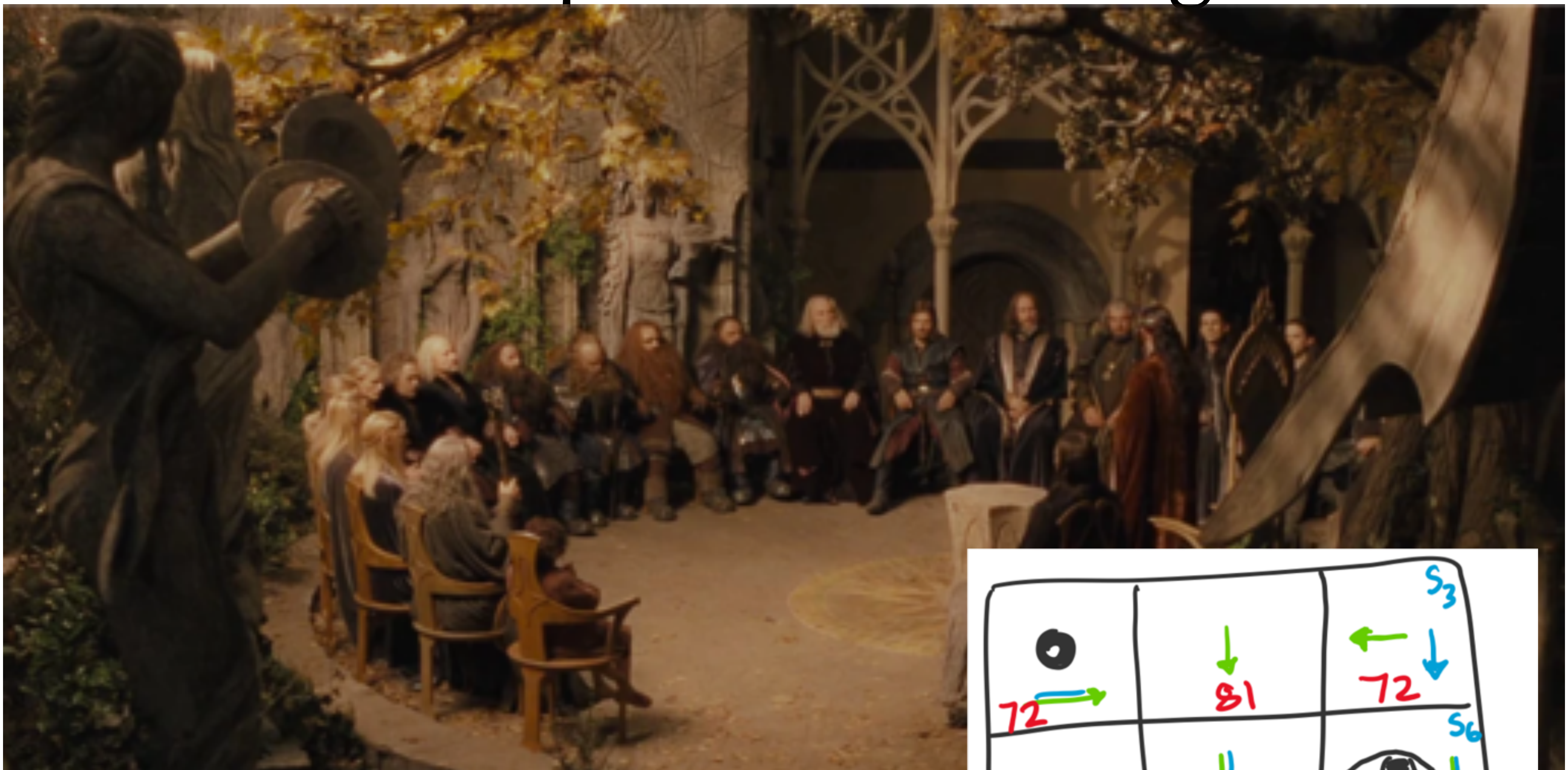
5 Graphical Models

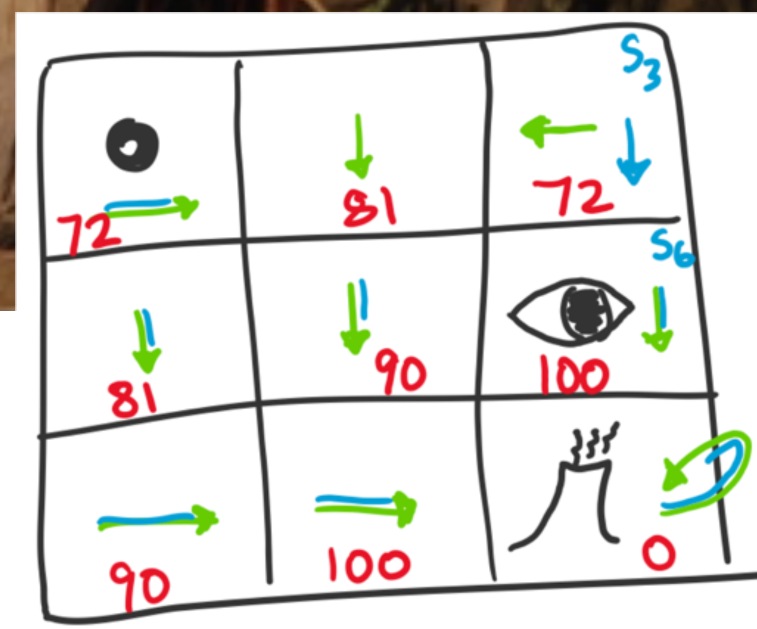
(f) [3 pts.] Give two reasons why the graphical models formalism is convenient when compared to learning a full joint distribution.

Gibbs Sampling



Example: Path Planning



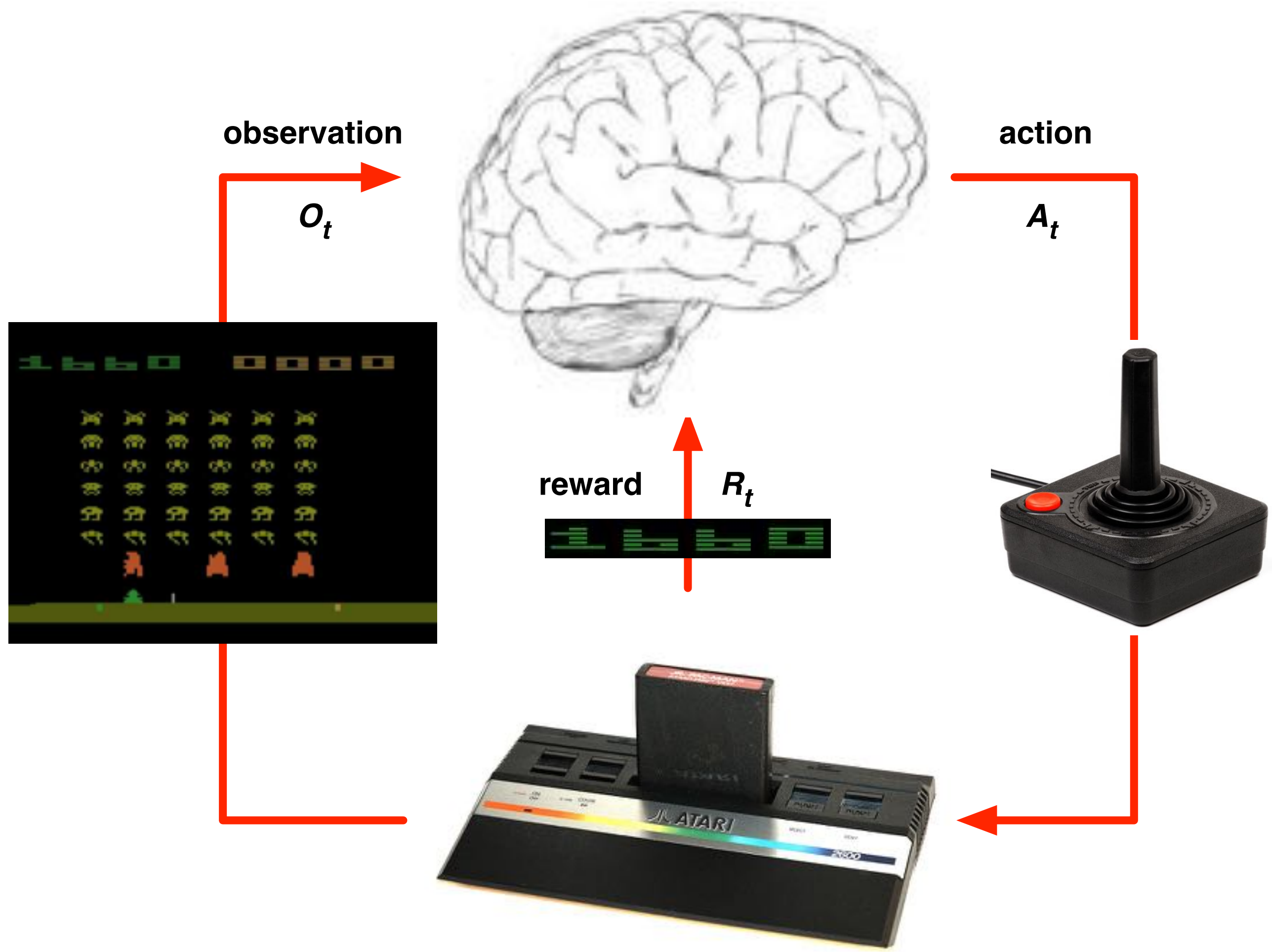


Today's lecture is brought you by the letter….

Q

Playing Atari with Deep RL

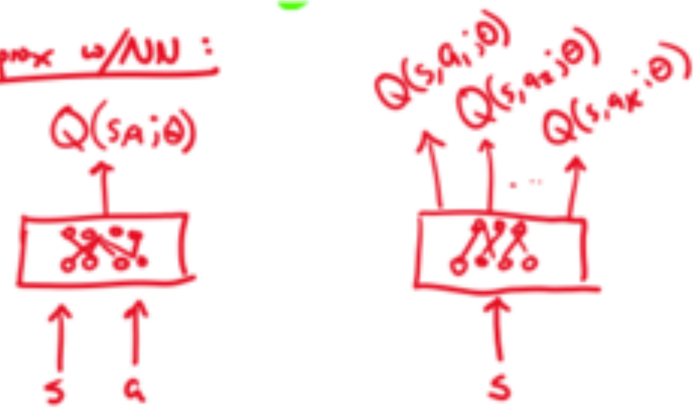
- Setup: RL system observes the pixels on the screen
- It receives rewards as the game score
- Actions decide how to move the joystick / buttons



Figures from David Silver (Intro RL lecture)

not-so-Deep Q-Learning

Approx w/NN:



$|A| = K = \#$ actions

Approx w/Linear Regression:

Represent state as a feature vector:

$$\vec{s}_t = [0,1,1,0,1,\ldots,1]^T \qquad |\vec{s}_t| \in \mathbb{R}^M$$

$M = \#$ features

$$Q(\vec{s},a;\Theta) = \vec{\Theta}_a^T \vec{s} \qquad \Theta = \begin{bmatrix} \text{---}\ \vec{\Theta}_1^T\ \text{---} \\ \text{---}\ \vec{\Theta}_2^T\ \text{---} \\ \vdots \\ \text{---}\ \vec{\Theta}_K^T\ \text{---} \end{bmatrix} \qquad \Theta \in \mathbb{R}^{K \times M}$$

$$\nabla_{\vec{\Theta}_b} Q(\vec{s},a;\Theta) = \ ?$$

$$= \begin{cases} \vec{s} & \text{if } a = b \\ \vec{0} & \text{if } a \neq b \end{cases}$$

$$= \vec{s}\, \mathbb{1}(a=b)$$

$$\nabla_{\Theta} Q(s,a;\Theta) = \begin{array}{cl} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ \text{---} & \vec{s}^T & \text{---} & \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} & \begin{matrix} \\ \\ \leftarrow a\text{th row} \\ \\ \\ \end{matrix} \end{array}$$

Sample Questions

7.1 Reinforcement Learning

A = calamity

3. (1 point) **Please select one statement that is true for reinforcement learning and supervised learning.**

B ◯ Reinforcement learning is a kind of supervised learning problem because you can treat the reward and next state as the label and each state, action pair as the training data.

C ✓ Reinforcement learning differs from supervised learning because it has a temporal structure in the learning process, whereas, in supervised learning, the prediction of a data point does not affect the data you would see in the future.

4. (1 point) **True or False:** Value iteration is better at balancing exploration and exploitation compared with policy iteration.

◯ True

✓ False

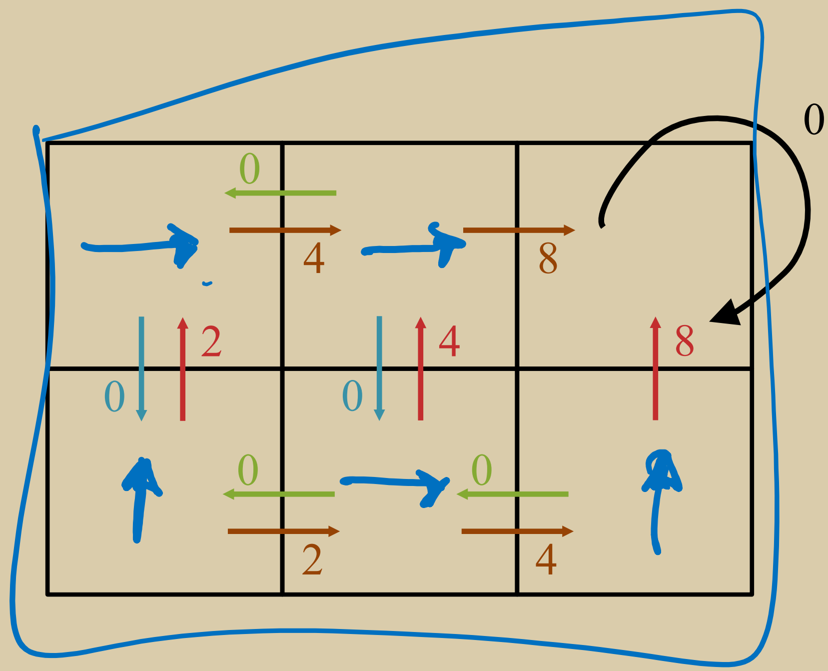
Sample Questions

7.1 Reinforcement Learning

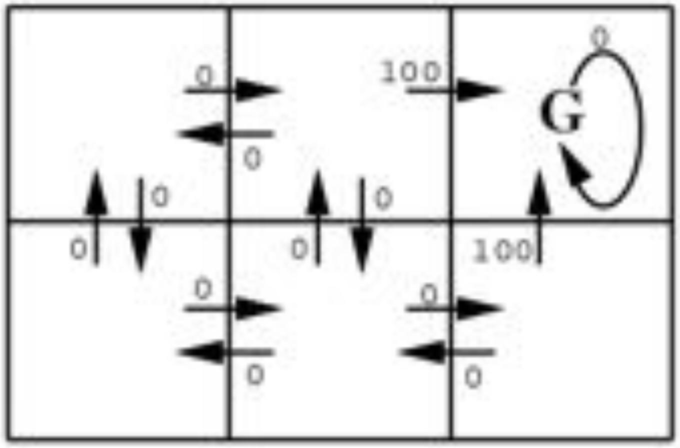
1. For the R(s,a) values shown on the arrows below, what is the corresponding optimal policy? Assume the discount factor is 0.1

2. For the R(s,a) values shown on the arrows below, which are the corresponding V*(s) values? Assume the discount factor is 0.1

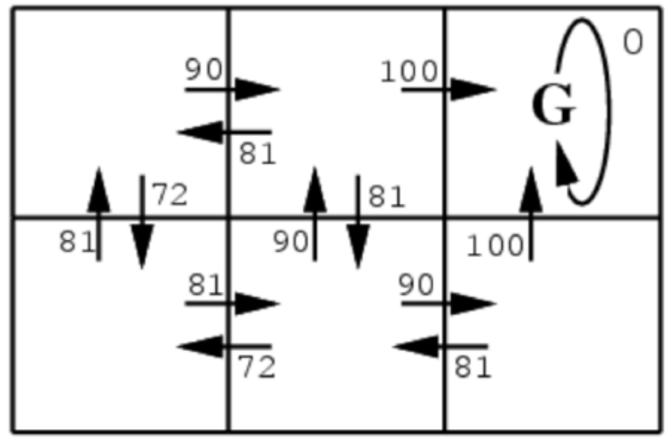
3. For the R(s,a) values shown on the arrows below, which are the corresponding Q*(s,a) values? Assume the discount factor is 0.1



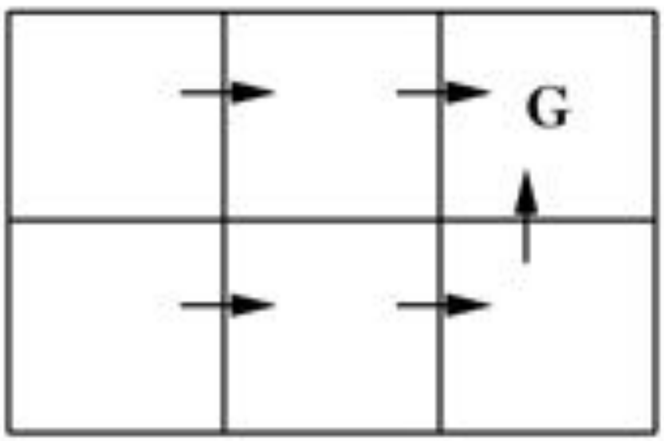
Example: Robot Localization



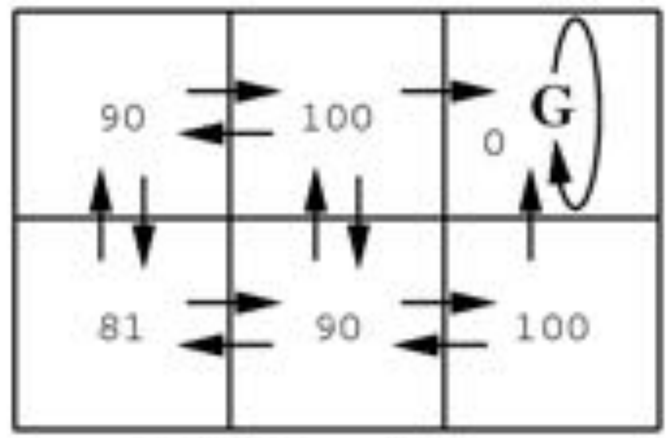
$r(s, a)$ (immediate reward) values



$Q(s, a)$ values



One optimal policy



$V^*(s)$ values

Figure from Tom Mitchell

K-Means Example: A Real-World Dataset

Example: K-Means

Example: K-Means

Q3

Samples Questions

2 K-Means Clustering

(a) [3 pts] We are given n data points, $x_1, ..., x_n$ and asked to cluster them using K-means. If we choose the value for k to optimize the objective function how many clusters will be used (i.e. what value of k will we choose)? **No justification required.**

(i) 1 A

(ii) 2 B

(iii) n C 60%

(iv) $\log(n)$ D 40%

E = calamity

aka. K-Means

Samples Questions

2.2 Lloyd's algorithm

Circle the image which depicts the cluster center positions after 1 iteration of Lloyd's algorithm.

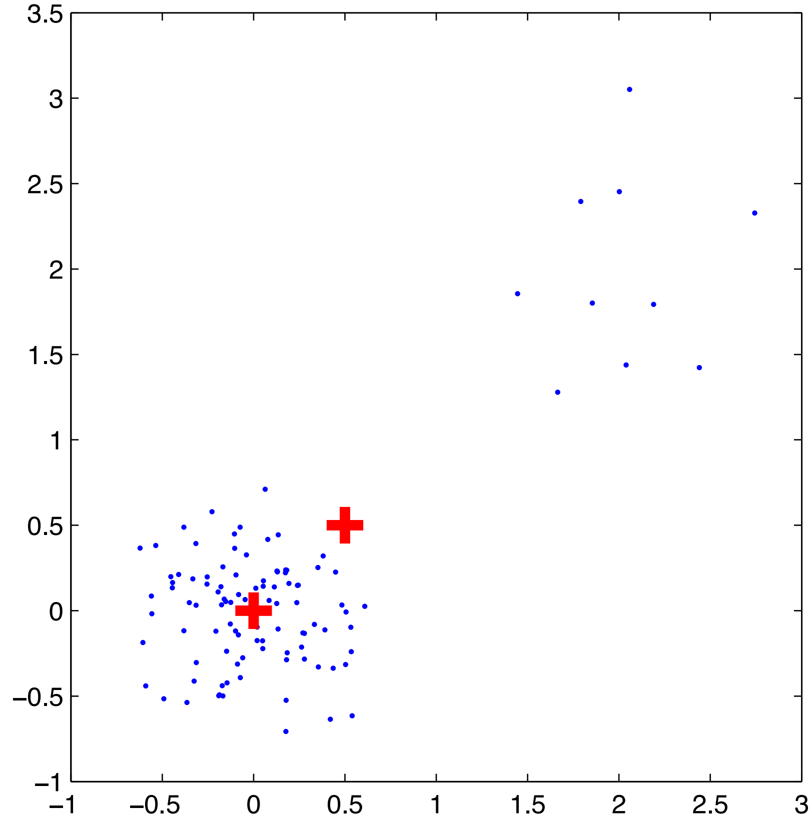


Figure 2: Initial data and cluster centers

Samples Questions

2.2 Lloyd's algorithm

Circle the image which depicts the cluster center positions after 1 iteration of Lloyd's algorithm.

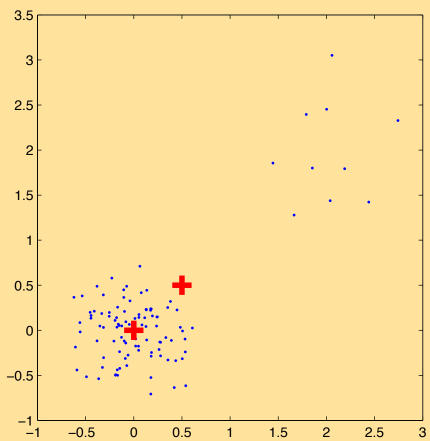
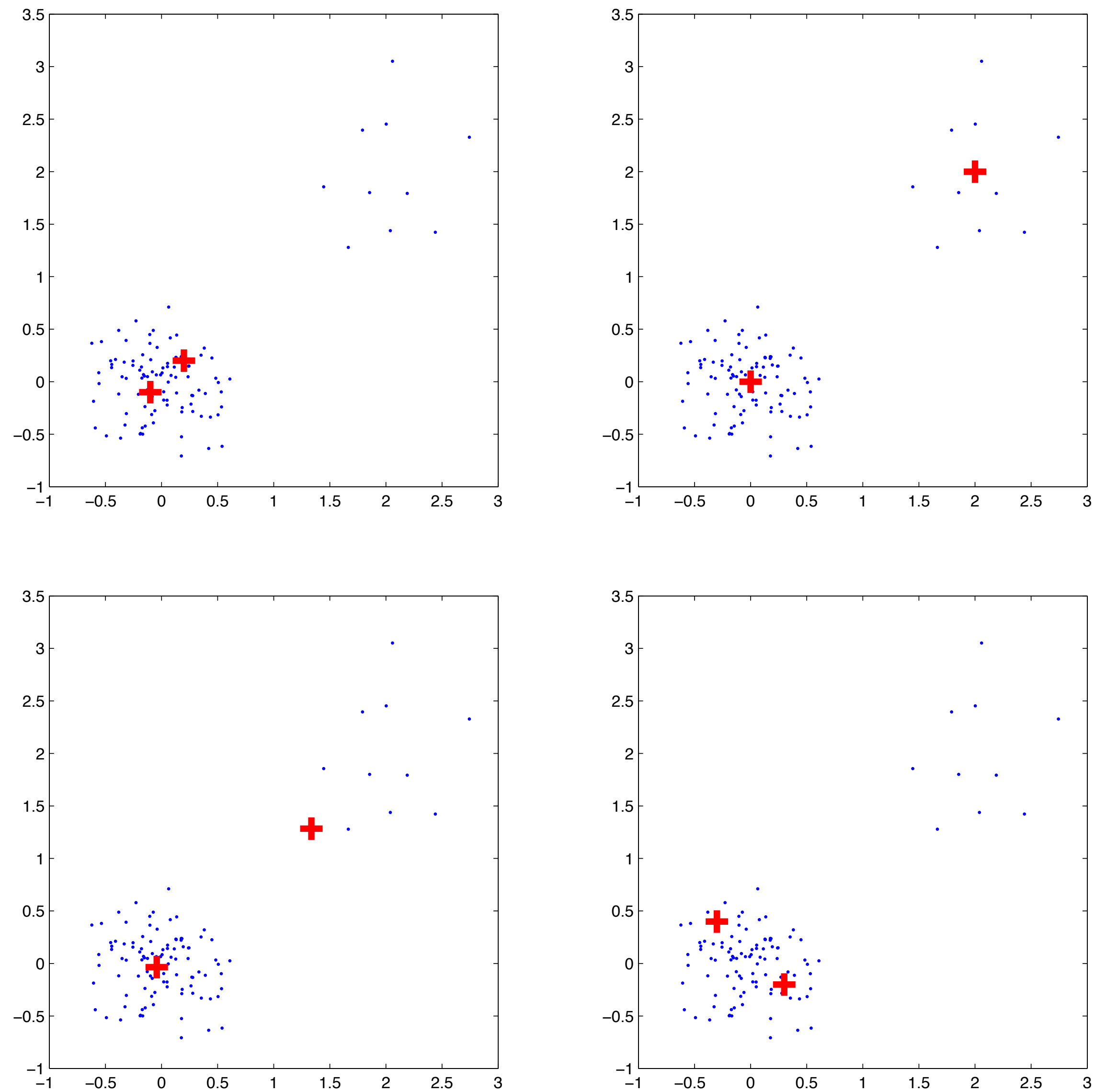


Figure 2: Initial data and cluster centers



High Dimension Data

Examples of high dimensional data:

– Brain Imaging Data (100s of MBs per scan)

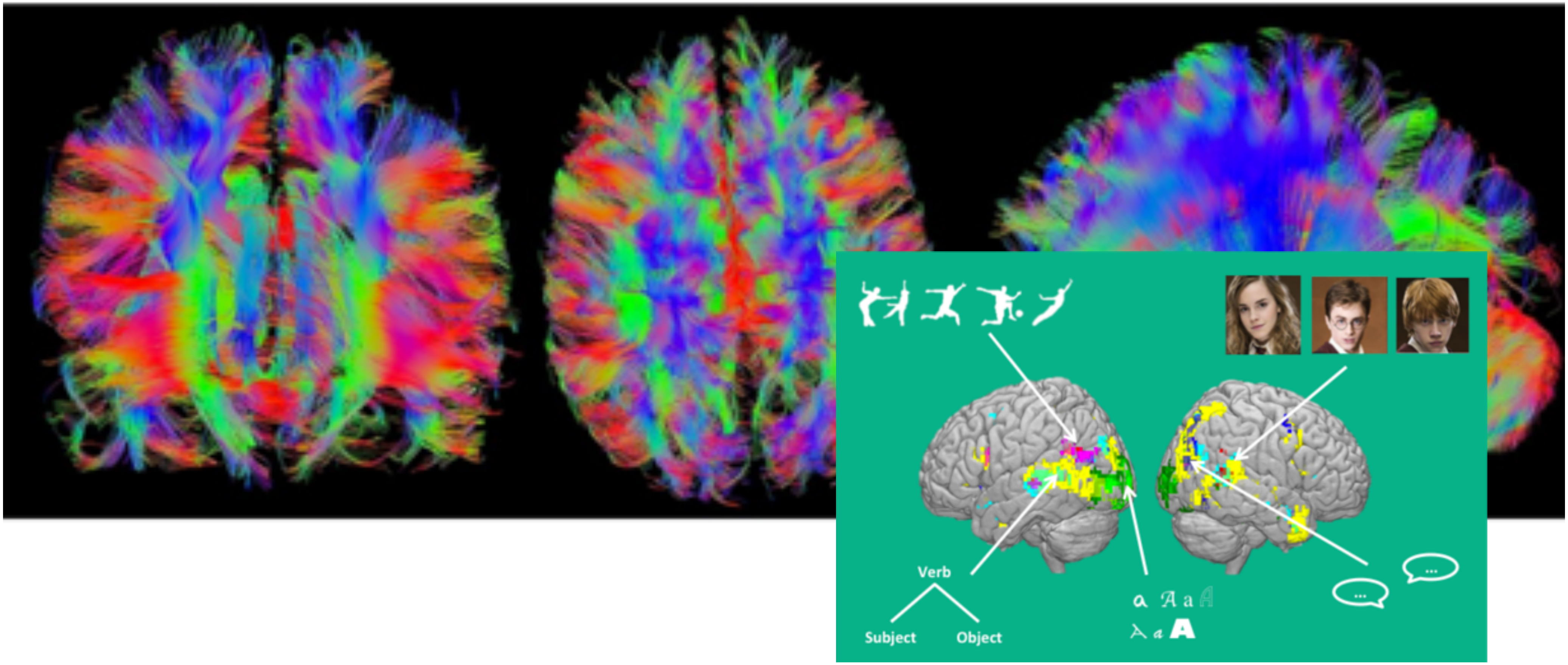


Image from (Wehbe et al., 2014)

Image from https://pixabay.com/en/brain-mrt-magnetic-resonance-imaging-1728449/

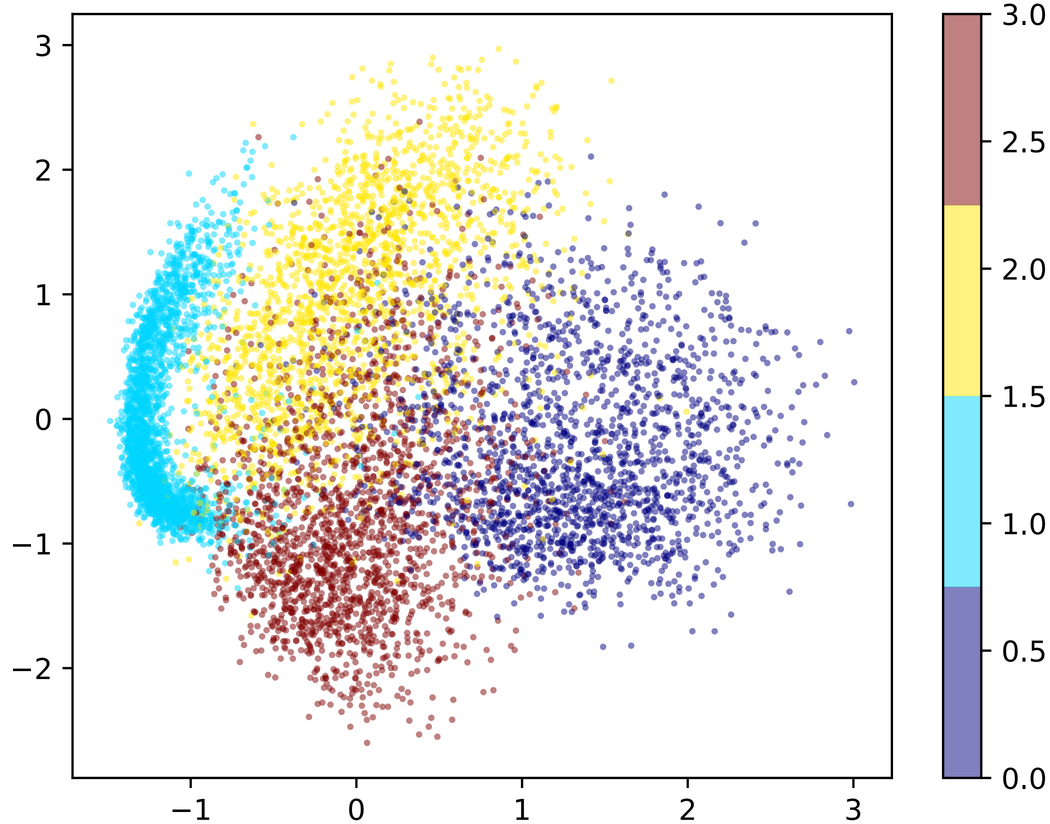
Shortcut Example

https://www.youtube.com/watch?v=MlJN9pEfPfE

Projecting MNIST digits

Task Setting:

1. Take 25x25 images of digits and project them down to 2 components
2. Plot the 2 dimensional points



Sample Questions

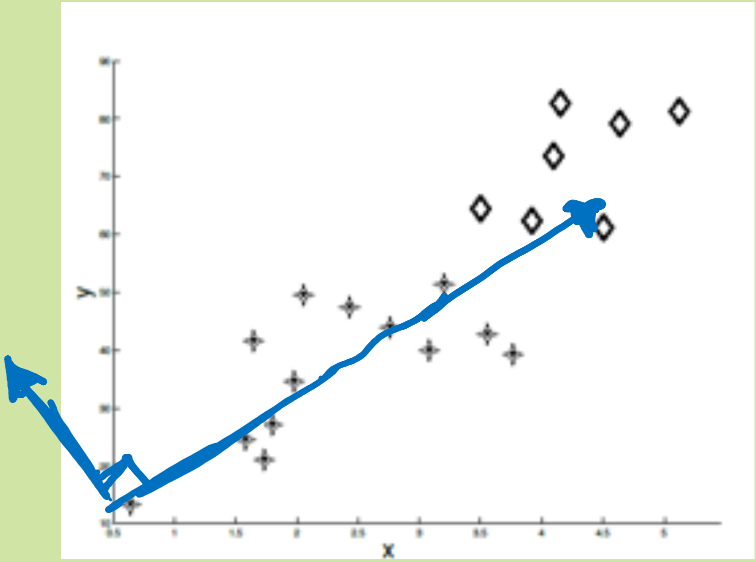
4 Principal Component Analysis [16 pts.]

(a) In the following plots, a train set of data points X belonging to two classes on $\mathbb{R}^2$ are given, where the original features are the coordinates (x, y). For each, answer the following questions:

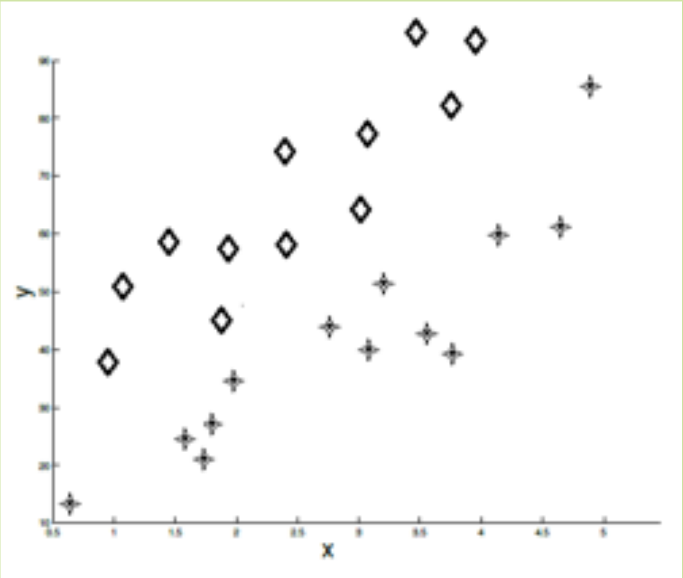
(i) [3 pt.] Draw all the principal components.

(ii) [6 pts.] Can we correctly classify this dataset by using a threshold function after projecting onto one of the principal components? If so, which principal component should we project onto? If not, explain in 1–2 sentences why it is not possible.

Dataset 1:



Dataset 2:



Sample Questions

4 Principal Component Analysis

(c) [2 pts.] Assume we apply PCA to a matrix $X \in \mathbb{R}^{n \times m}$ and obtain a set of PCA features, $Z \in \mathbb{R}^{m \times n}$. We divide this set into two, Z_1 and Z_2. The first set, Z_1, corresponds to the top principal components. The second set, Z_2, corresponds to the remaining principal components. Which is more common in the training data:

A: a point with large feature values in Z_1 and small feature values in Z_2

B: a point with large feature values in Z_2 and small feature values in Z_1

Sample Questions

4 Principal Component Analysis

(i) **T or F** The goal of PCA is to interpret the underlying structure of the data in terms of the principal components that are best at predicting the output variable.

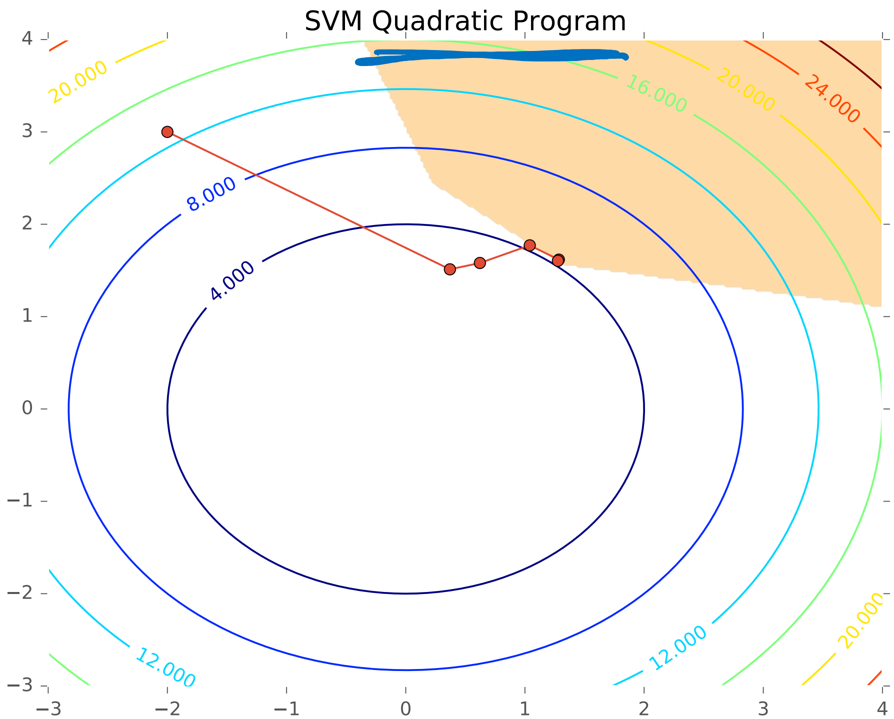
(ii) **T or F** The output of PCA is a new representation of the data that is always of lower dimensionality than the original feature representation.

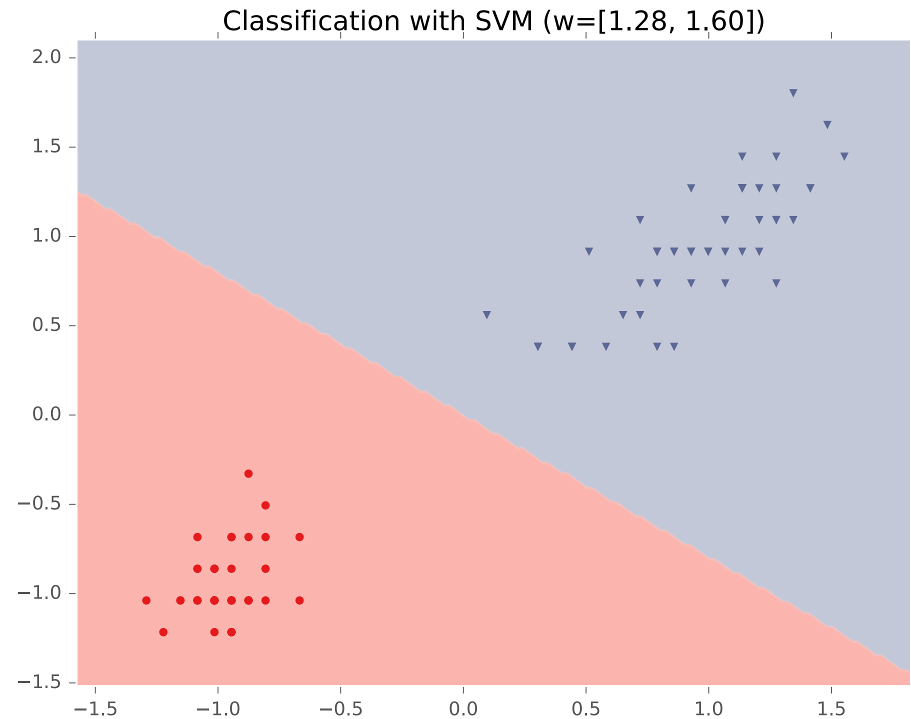
(iii) **T or F** Subsequent principal components are always orthogonal to each other.

SVM Example: Building Walls

https://www.facebook.com/Mondobloxx/

SVM QP


SVM Quadratic Program
20.000
16.000
24.000
8.000
4.000
12.000


Classification with SVM (w=[1.28, 1.60])

Soft-Margin SVM

Hard-margin SVM (Primal)

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t. } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1, \quad \forall i = 1, \dots, N$$

Hard-margin SVM (Lagrangian Dual)

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y^{(i)} y^{(j)} \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)}$$

$$\text{s.t. } \alpha_i \geq 0, \quad \forall i = 1, \dots, N$$

$$\sum_{i=1}^{N} \alpha_i y^{(i)} = 0$$

Soft-margin SVM (Primal)

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\left(\sum_{i=1}^{N} e_i\right)$$

$$\text{s.t. } y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - e_i, \quad \forall i = 1, \dots, N$$

$$e_i \geq 0, \quad \forall i = 1, \dots, N$$

Soft-margin SVM (Lagrangian Dual)

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y^{(i)} y^{(j)} \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)}$$

$$\text{s.t. } 0 \leq \alpha_i \leq C, \quad \forall i = 1, \dots, N$$

$$\sum_{i=1}^{N} \alpha_i y^{(i)} = 0$$

Sample Questions

(c) [4 pts.] **Extra Credit:** Consider the dataset in Fig. 4. Under the SVM formulation in section 4.2(a),

(1) Draw the decision boundary on the graph.

(2) What is the size of the margin?

(3) Circle all the support vectors on the graph.

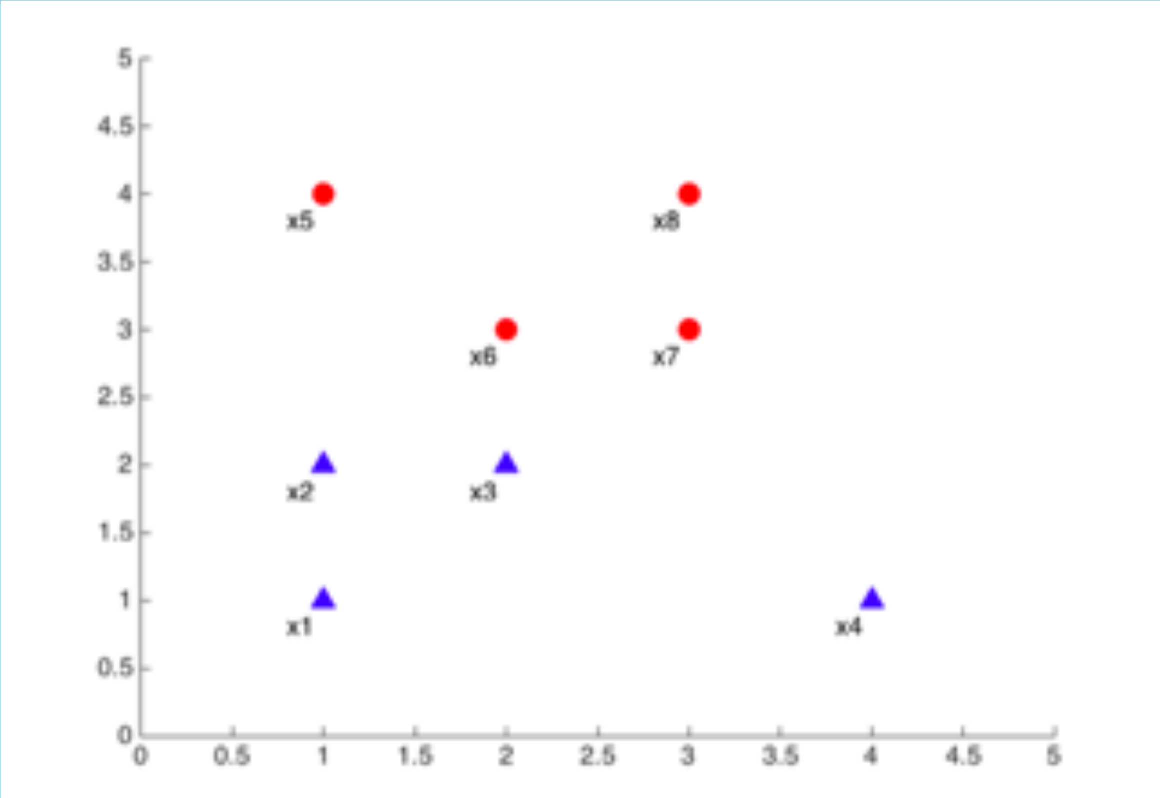


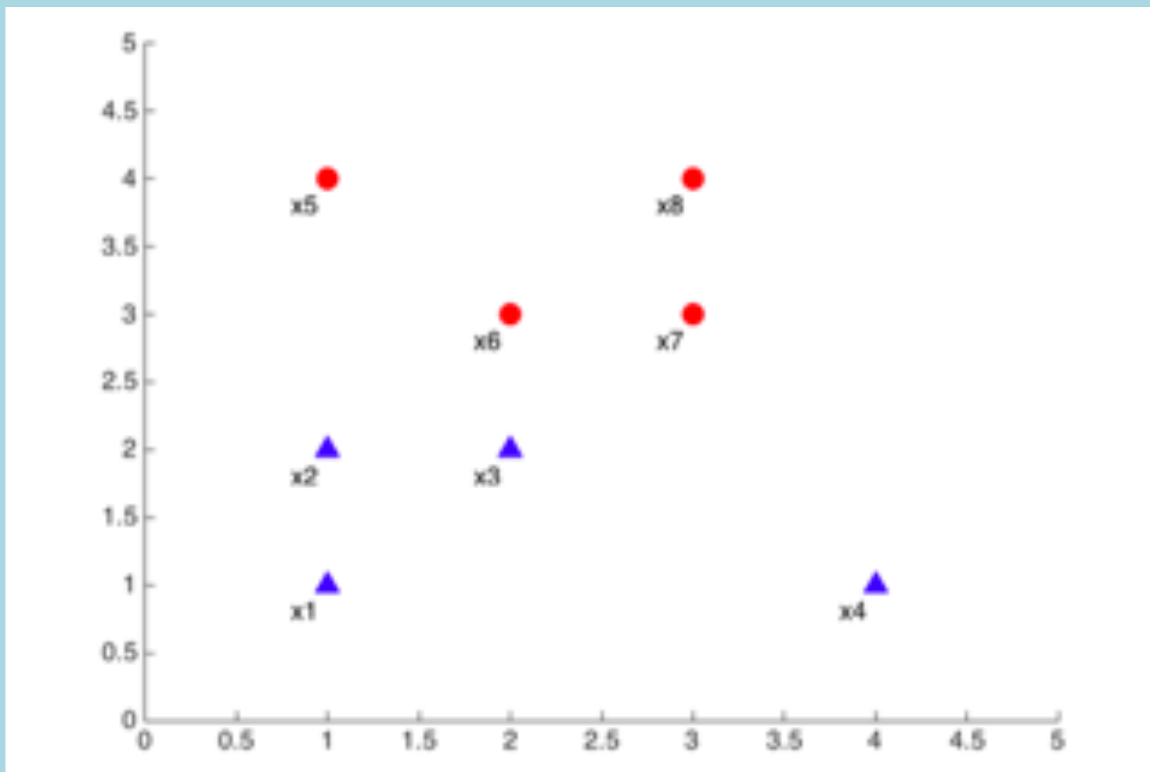
Figure 4: SVM toy dataset

Sample Questions

4.2 Multiple Choice

(a) [3 pt.] If the data is linearly separable, SVM minimizes $\|w\|^2$ subject to the constraints $\forall i, y_i w \cdot x_i \geq 1$. In the linearly separable case, which of the following may happen to the decision boundary if one of the training samples is removed? **Circle all that apply.**

- Shifts toward the point removed
- Shifts away from the point removed
- Does not change



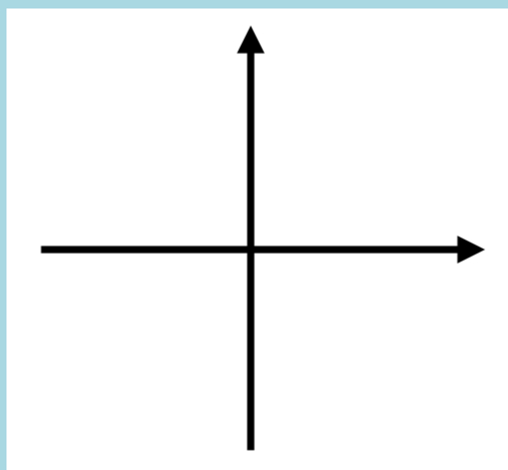
Sample Questions

3. **[Extra Credit: 3 pts.]** One formulation of soft-margin SVM optimization problem is:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t. } y_i(\mathbf{w}^\top x_i) \geq 1 - \xi_i \quad \forall i = 1, ..., N$$
$$\xi_i \geq 0 \quad \forall i = 1, ..., N$$
$$C \geq 0$$

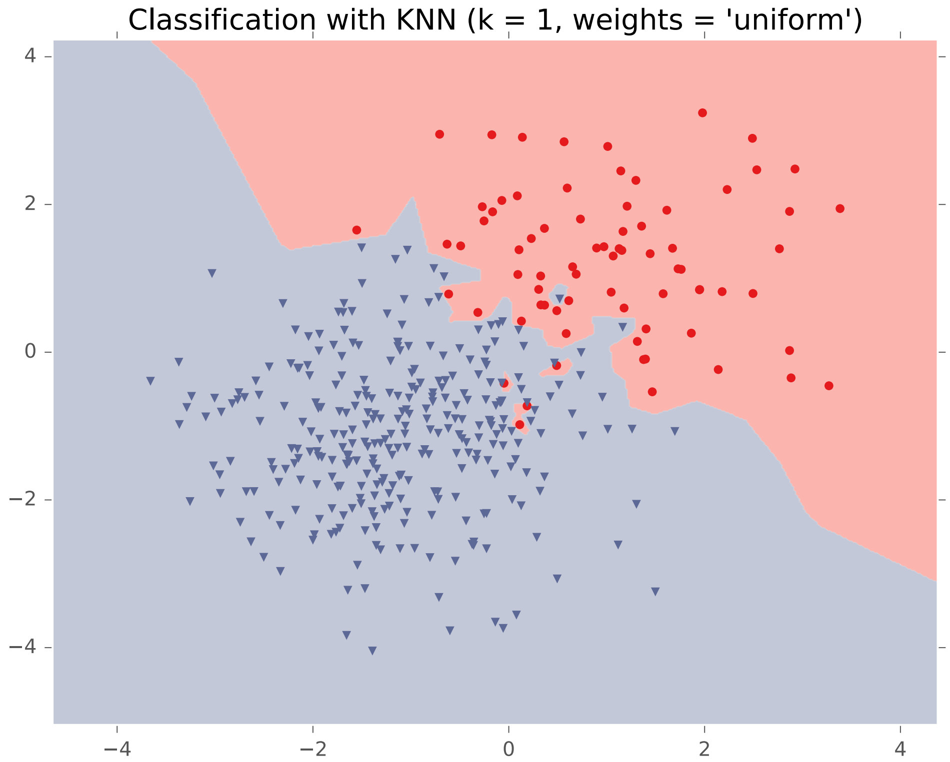
where (x_i, y_i) are training samples and $\mathbf{w}$ defines a linear decision boundary.

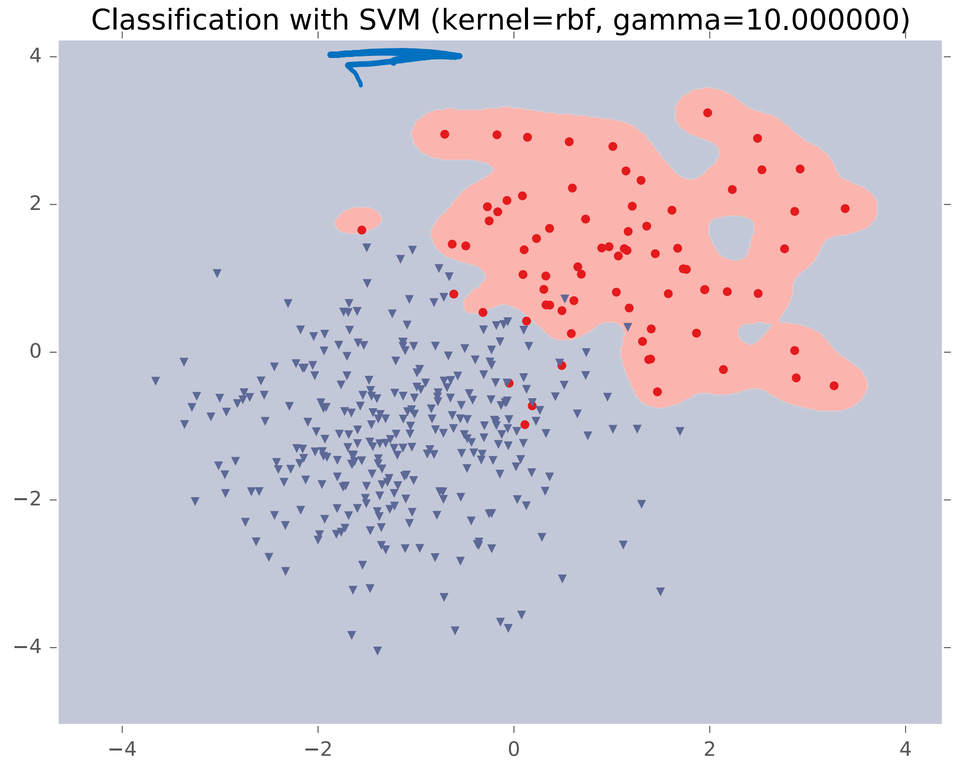
Derive a formula for ξ_i when the objective function achieves its minimum (No steps necessary). Note it is a function of $y_i\mathbf{w}^\top x_i$. Sketch a plot of ξ_i with $y_i\mathbf{w}^\top x_i$ on the x-axis and value of ξ_i on the y-axis. What is the name of this function?



RBF Kernel Example

KNN vs. SVM





RBF Kernel: $K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \exp(-\gamma||\mathbf{x}^{(i)} - \mathbf{x}^{(j)}||_2^2)$

Sample Questions

4.3 Analysis

(a) [4 pts.] In one or two sentences, describe the benefit of using the Kernel trick.

(b) [4 pt.] The concept of margin is essential in both SVM and Perceptron. Describe why a large margin separator is desirable for classification.

(e) [2 pts.] **T or F**: The function $K(\mathbf{x}, \mathbf{z}) = -2\mathbf{x}^T\mathbf{z}$ is a valid kernel function.

Recommender Systems

NETFLIX

Netflix Prize

COMPLETED

Home | Rules | Leaderboard | Update

Leaderboard

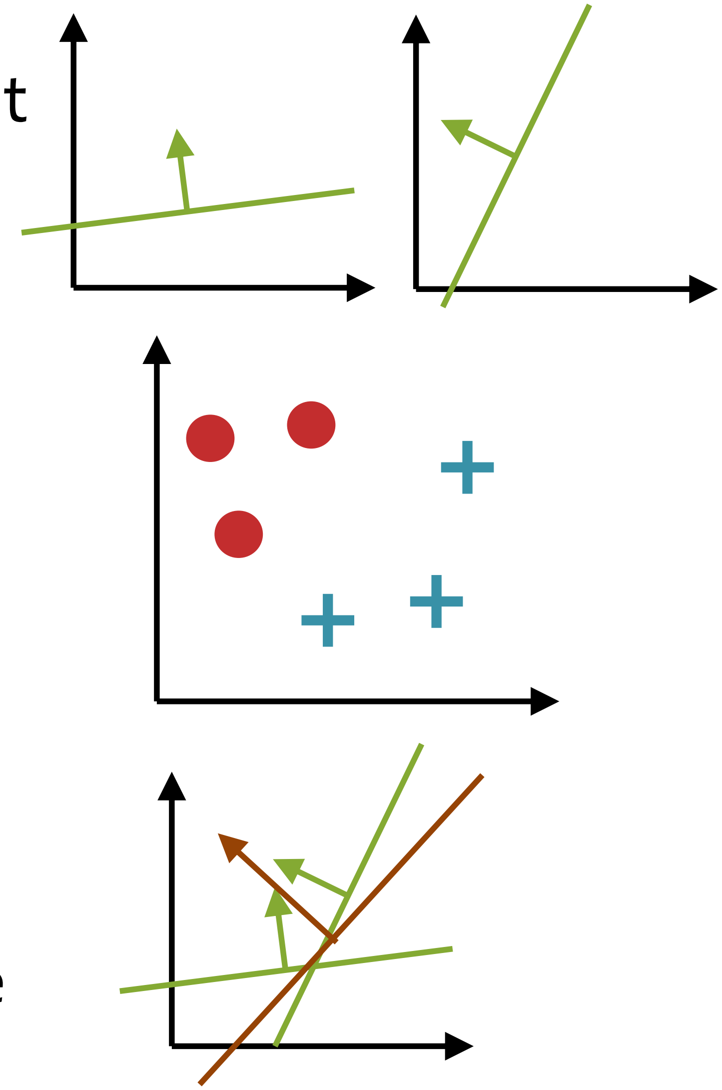
Showing Test Score. Click here to show quiz score

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

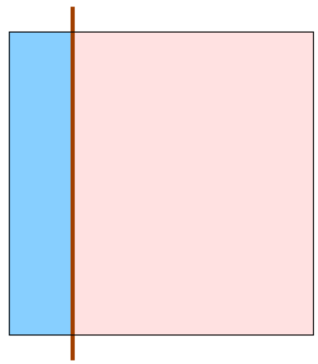
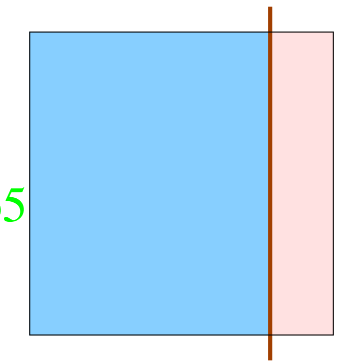
Weighted Majority Algorithm

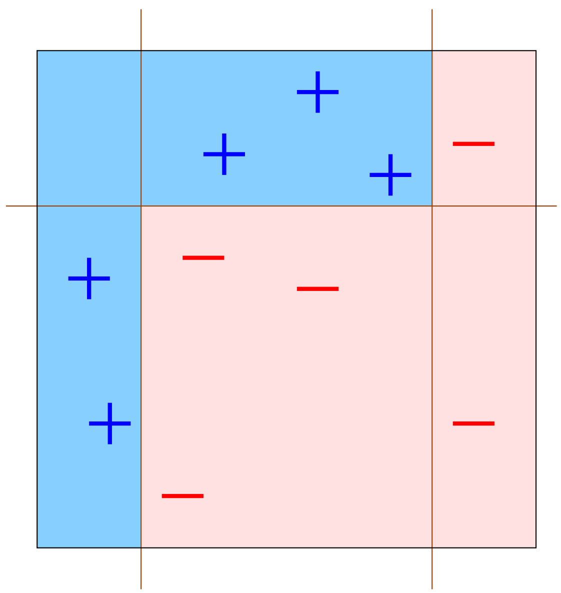
(Littlestone & Warmuth, 1994)

- **Given:** pool *A* of binary classifiers (that you know nothing about)
- **Data:** stream of examples (i.e. online learning setting)
- **Goal:** design a new learner that uses the predictions of the pool to make new predictions
- **Algorithm:**
  - Initially weight all classifiers equally
  - Receive a training example and predict the (weighted) majority vote of the classifiers in the pool
  - Down-weight classifiers that contribute to a mistake by a factor of β



AdaBoost: Toy Example

$H_{\text{final}} = \text{sign}\Big(0.42$  $+ 0.65$  $+ 0.92$  $\Big)$

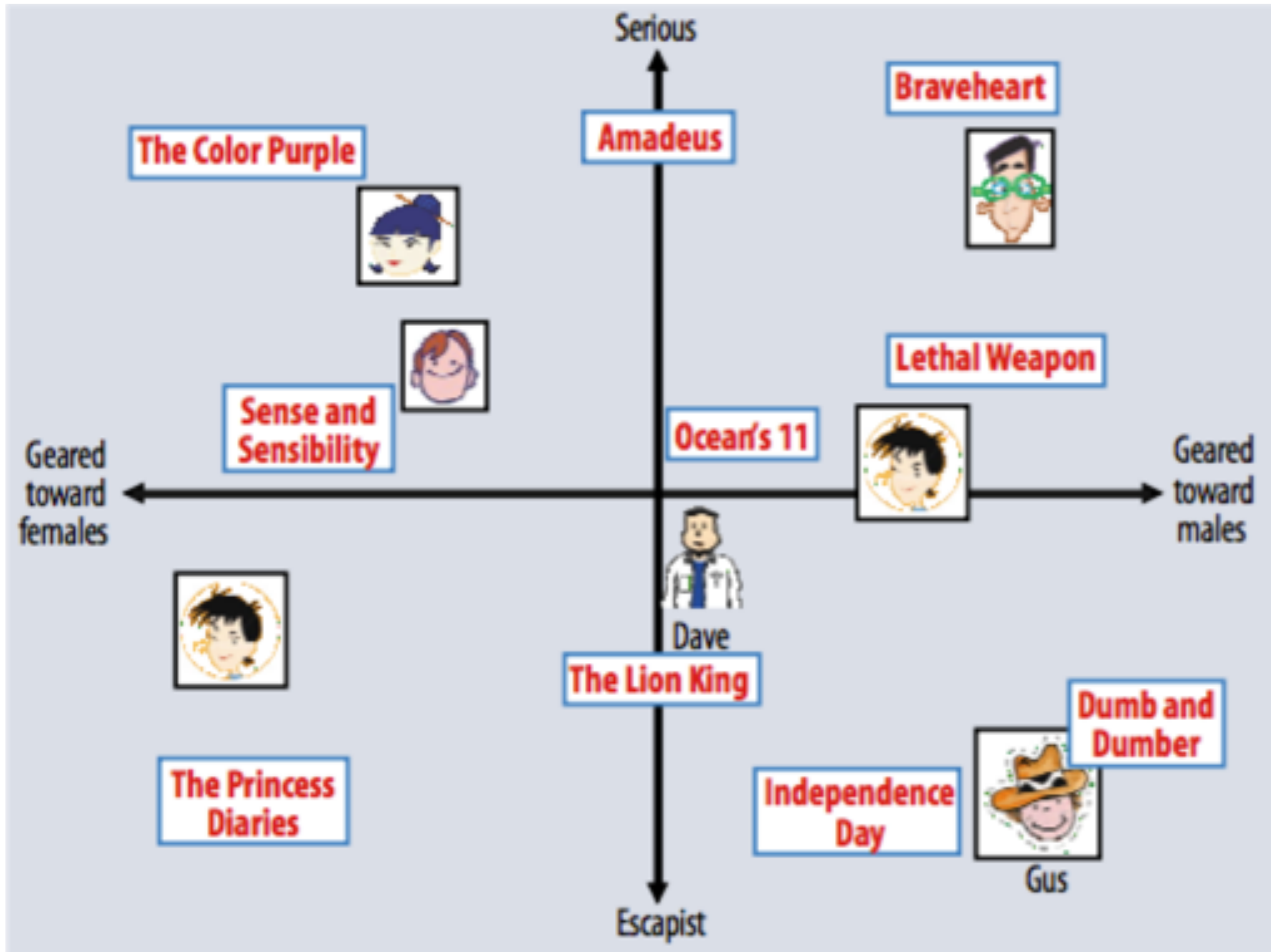
$=$ 

Slide from Schapire NIPS Tutorial

Two Types of Collaborative Filtering

2. Latent Factor Methods

- Assume that both movies and users live in some **low-dimensional space** describing their properties
- **Recommend** a movie based on its **proximity** to the user in the latent space
- **Example Algorithm:** Matrix Factorization



Figures from Koren et al. (2009)

Crowdsourcing Exam Questions

In-Class Exercise

1. Select one of lecture-level learning objectives
   http://mlcourse.org/slides/10601-objectives.pdf
2. Write a question that assesses that objective
3. Adjust to avoid 'trivia style' question

Answer Here:

The Big Picture

MACHINE LEARNING

Learning Paradigms

Paradigm	Data	
Supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$	$\mathbf{x} \sim p^*(\cdot)$ and $y = c^*(\cdot)$
↪ Regression	$y^{(i)} \in \mathbb{R}$	
↪ Classification	$y^{(i)} \in \{1, \ldots, K\}$	
↪ Binary classification	$y^{(i)} \in \{+1, -1\}$	
↪ Structured Prediction	$\mathbf{y}^{(i)}$ is a vector	

Learning Paradigms

Paradigm	Data	
Supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$	$\mathbf{x} \sim p^*(\cdot)$ and $y = c^*(\cdot)$
$\hookrightarrow$ Regression	$y^{(i)} \in \mathbb{R}$	
$\hookrightarrow$ Classification	$y^{(i)} \in \{1, \ldots, K\}$	
$\hookrightarrow$ Binary classification	$y^{(i)} \in \{+1, -1\}$	
$\hookrightarrow$ Structured Prediction	$\mathbf{y}^{(i)}$ is a vector	
Unsupervised	$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$	$\mathbf{x} \sim p^*(\cdot)$

Learning Paradigms

Paradigm	**Data**
Supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N} \qquad \mathbf{x} \sim p^*(\cdot)$ and $y = c^*(\cdot)$
↪ Regression	$y^{(i)} \in \mathbb{R}$
↪ Classification	$y^{(i)} \in \{1, \ldots, K\}$
↪ Binary classification	$y^{(i)} \in \{+1, -1\}$
↪ Structured Prediction	$\mathbf{y}^{(i)}$ is a vector
Unsupervised	$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N} \qquad \mathbf{x} \sim p^*(\cdot)$
Semi-supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N_1} \cup \{\mathbf{x}^{(j)}\}_{j=1}^{N_2}$

Learning Paradigms

Paradigm	Data
Supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N \qquad \mathbf{x} \sim p^*(\cdot)$ and $y = c^*(\cdot)$
$\hookrightarrow$ Regression	$y^{(i)} \in \mathbb{R}$
$\hookrightarrow$ Classification	$y^{(i)} \in \{1, \ldots, K\}$
$\hookrightarrow$ Binary classification	$y^{(i)} \in \{+1, -1\}$
$\hookrightarrow$ Structured Prediction	$\mathbf{y}^{(i)}$ is a vector
Unsupervised	$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N \qquad \mathbf{x} \sim p^*(\cdot)$
Semi-supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N_1} \cup \{\mathbf{x}^{(j)}\}_{j=1}^{N_2}$
Online	$\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), (\mathbf{x}^{(3)}, y^{(3)}), \ldots\}$

Learning Paradigms

Paradigm	**Data**
Supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N \qquad \mathbf{x} \sim p^*(\cdot)$ and $y = c^*(\cdot)$
$\hookrightarrow$ Regression	$y^{(i)} \in \mathbb{R}$
$\hookrightarrow$ Classification	$y^{(i)} \in \{1, \ldots, K\}$
$\hookrightarrow$ Binary classification	$y^{(i)} \in \{+1, -1\}$
$\hookrightarrow$ Structured Prediction	$\mathbf{y}^{(i)}$ is a vector
Unsupervised	$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N \qquad \mathbf{x} \sim p^*(\cdot)$
Semi-supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N_1} \cup \{\mathbf{x}^{(j)}\}_{j=1}^{N_2}$
Online	$\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), (\mathbf{x}^{(3)}, y^{(3)}), \ldots\}$
Active Learning	$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ and can query $y^{(i)} = c^*(\cdot)$ at a cost

Learning Paradigms

Paradigm	Data
Supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N \quad \mathbf{x} \sim p^*(\cdot)$ and $y = c^*(\cdot)$
$\hookrightarrow$ Regression	$y^{(i)} \in \mathbb{R}$
$\hookrightarrow$ Classification	$y^{(i)} \in \{1, \ldots, K\}$
$\hookrightarrow$ Binary classification	$y^{(i)} \in \{+1, -1\}$
$\hookrightarrow$ Structured Prediction	$\mathbf{y}^{(i)}$ is a vector
Unsupervised	$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N \quad \mathbf{x} \sim p^*(\cdot)$
Semi-supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N_1} \cup \{\mathbf{x}^{(j)}\}_{j=1}^{N_2}$
Online	$\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), (\mathbf{x}^{(3)}, y^{(3)}), \ldots\}$
Active Learning	$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ and can query $y^{(i)} = c^*(\cdot)$ at a cost
Imitation Learning	$\mathcal{D} = \{(s^{(1)}, a^{(1)}), (s^{(2)}, a^{(2)}), \ldots\}$

Learning Paradigms

Paradigm	**Data**
Supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$ $\quad \mathbf{x} \sim p^*(\cdot)$ and $y = c^*(\cdot)$
$\hookrightarrow$ Regression	$y^{(i)} \in \mathbb{R}$
$\hookrightarrow$ Classification	$y^{(i)} \in \{1, \ldots, K\}$
$\hookrightarrow$ Binary classification	$y^{(i)} \in \{+1, -1\}$
$\hookrightarrow$ Structured Prediction	$\mathbf{y}^{(i)}$ is a vector
Unsupervised	$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ $\quad \mathbf{x} \sim p^*(\cdot)$
Semi-supervised	$\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N_1} \cup \{\mathbf{x}^{(j)}\}_{j=1}^{N_2}$
Online	$\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), (\mathbf{x}^{(3)}, y^{(3)}), \ldots\}$
Active Learning	$\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ and can query $y^{(i)} = c^*(\cdot)$ at a cost
Imitation Learning	$\mathcal{D} = \{(s^{(1)}, a^{(1)}), (s^{(2)}, a^{(2)}), \ldots\}$
Reinforcement Learning	$\mathcal{D} = \{(s^{(1)}, a^{(1)}, r^{(1)}), (s^{(2)}, a^{(2)}, r^{(2)}), \ldots\}$

Machine Learning: The Big Picture

Whiteboard

- **Decision Rules / Models** (probabilistic generative, probabilistic discriminative, perceptron, SVM, regression, MDP, graphical models)
- **Objective Functions** (likelihood, conditional likelihood, hinge loss, mean squared error)
- **Regularization** (L1, L2, priors for MAP)
- **Update Rules** (SGD, perceptron)
- **Nonlinear Features** (preprocessing, kernel trick)

ML Big Picture

Learning Paradigms:

What data is available and when? What form of prediction?

- supervised learning
- unsupervised learning
- semi-supervised learning
- reinforcement learning
- active learning
- imitation learning
- domain adaptation
- online learning
- density estimation
- recommender systems
- feature learning
- manifold learning
- dimensionality reduction
- ensemble learning
- distant supervision
- hyperparameter optimization

Problem Formulation:

What is the structure of our output prediction?

boolean	Binary Classification
categorical	Multiclass Classification
ordinal	Ordinal Classification
real	Regression
ordering	Ranking
multiple discrete	Structured Prediction
multiple continuous	(e.g. dynamical systems)
both discrete & cont.	(e.g. mixed graphical models)

Application Areas

Key challenges?

NLP, Speech, Computer Vision, Robotics, Medicine, Search

Theoretical Foundations:

What principles guide learning?

- ❑ probabilistic
- ❑ information theoretic
- ❑ evolutionary search
- ❑ ML as optimization

Facets of Building ML Systems:

How to build systems that are robust, efficient, adaptive, effective?

1. *Data prep*
2. *Model selection*
3. *Training (optimization / search)*
4. *Hyperparameter tuning on validation data*
5. *(Blind) Assessment on test data*

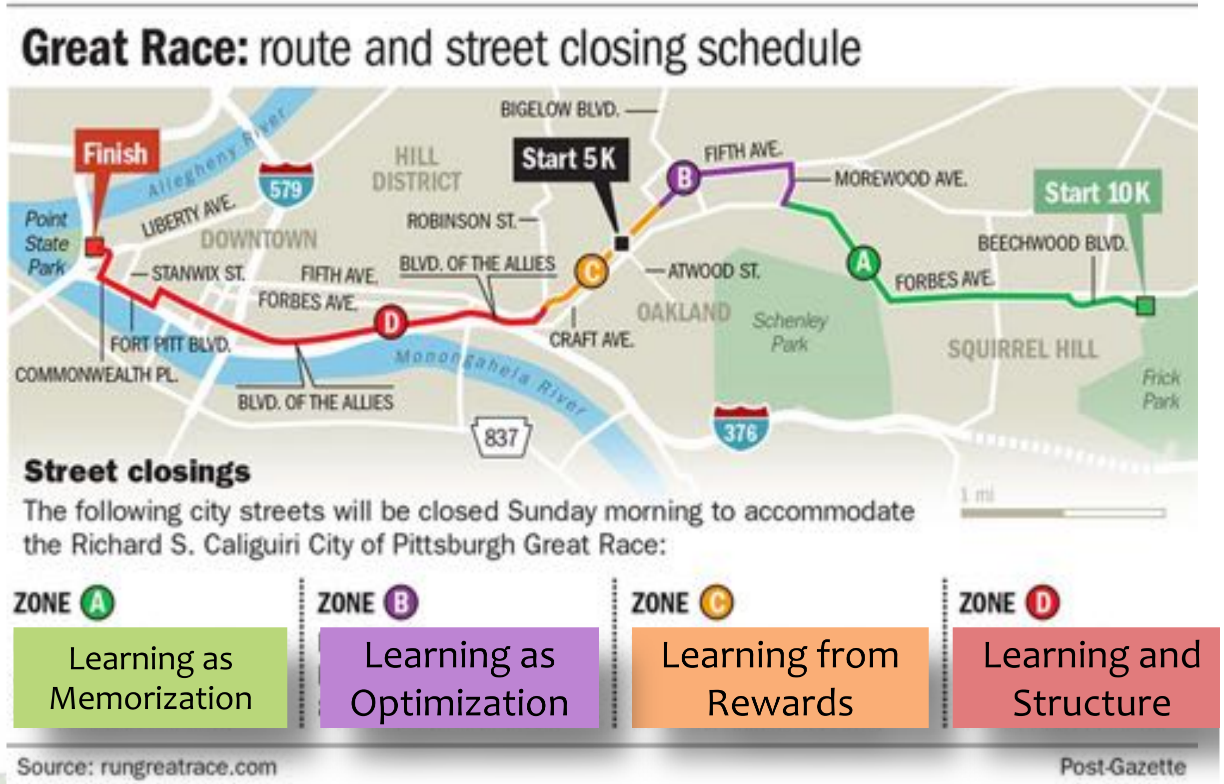
Big Ideas in ML:

Which are the ideas driving development of the field?

- *inductive bias*
- *generalization / overfitting*
- *bias-variance decomposition*
- *generative vs. discriminative*
- *deep nets, graphical models*
- *PAC learning*
- *distant rewards*

A new **combined** course…

…with the best (uphill climbs) from both



Course Level Objectives

You should be able to…

1. Implement and analyze existing learning algorithms, including well-studied methods for classification, regression, structured prediction, clustering, and representation learning
2. Integrate multiple facets of practical machine learning in a single system: data preprocessing, learning, regularization and model selection
3. Describe the the formal properties of models and algorithms for learning and explain the practical implications of those results
4. Compare and contrast different paradigms for learning (supervised, unsupervised, etc.)
5. Design experiments to evaluate and compare different machine learning techniques on real-world problems
6. Employ probability, statistics, calculus, linear algebra, and optimization in order to develop new predictive models or learning methods
7. Given a description of a ML technique, analyze it to identify (1) the expressive power of the formalism; (2) the inductive bias implicit in the algorithm; (3) the size and complexity of the search space; (4) the computational properties of the algorithm: (5) any guarantees (or lack thereof) regarding termination, convergence, correctness, accuracy or generalization power.

Q&A