10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Final Exam Review

Matt Gormley
Lecture 29
May 1, 2019

Reminders

- Homework 9: Learning Paradigms
  - Out: Wed, Apr 24
  - Due: Wed, May 1 at 11:59pm
  - Can only be submitted up to 3 days late, so we can return grades before final exam

- Today's In-Class Poll
  - http://p28.mlcourse.org

EXAM LOGISTICS

Final Exam

- **Time / Location**
  - **Time:** Evening Exam
    Mon, May 1 at 1:00pm – 4:00pm
  - **Room:** We will contact each student individually with **your room assignment.** The rooms are **not** based on section.
  - **Seats:** There will be **assigned seats.** Please arrive early.
  - Please watch Piazza carefully for announcements regarding room / seat assignments.
- **Logistics**
  - Format of questions:
    - Multiple choice
    - True / False (with justification)
    - Derivations
    - Short answers
    - Interpreting figures
    - Implementing algorithms on paper
  - No electronic devices
  - You are allowed to **bring** one 8½ x 11 sheet of notes (front and back)

Final Exam

- **How to Prepare**
  - Attend (or watch) this final exam review session
  - Review prior year's exams and solutions
    - We already posted these for the midterm
    - Disclaimer: This year's 10-601 is not the same as prior offerings, so review both midterm and final
  - Solve the "Final Exam **Worksheet 1**" and "Final Exam **Worksheet 2**" problems
    - Note: We'll release Worksheet 2 in time for the Recitation on Friday
  - 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
  - Attend the **Final Exam Recitation** (Friday)

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

Final Exam

- **Exam Contents**
  - ~30% of material comes from topics covered **before** Midterm Exam 2
  - ~70% of material comes from topics covered **after** Midterm Exam 2

Topics for Midterm 1

- Foundations
  - Probability, Linear Algebra, Geometry, Calculus
  - Optimization
- Important Concepts
  - Overfitting
  - Experimental Design

- Classification
  - Decision Tree
  - KNN
  - Perceptron
- Regression
  - Linear Regression

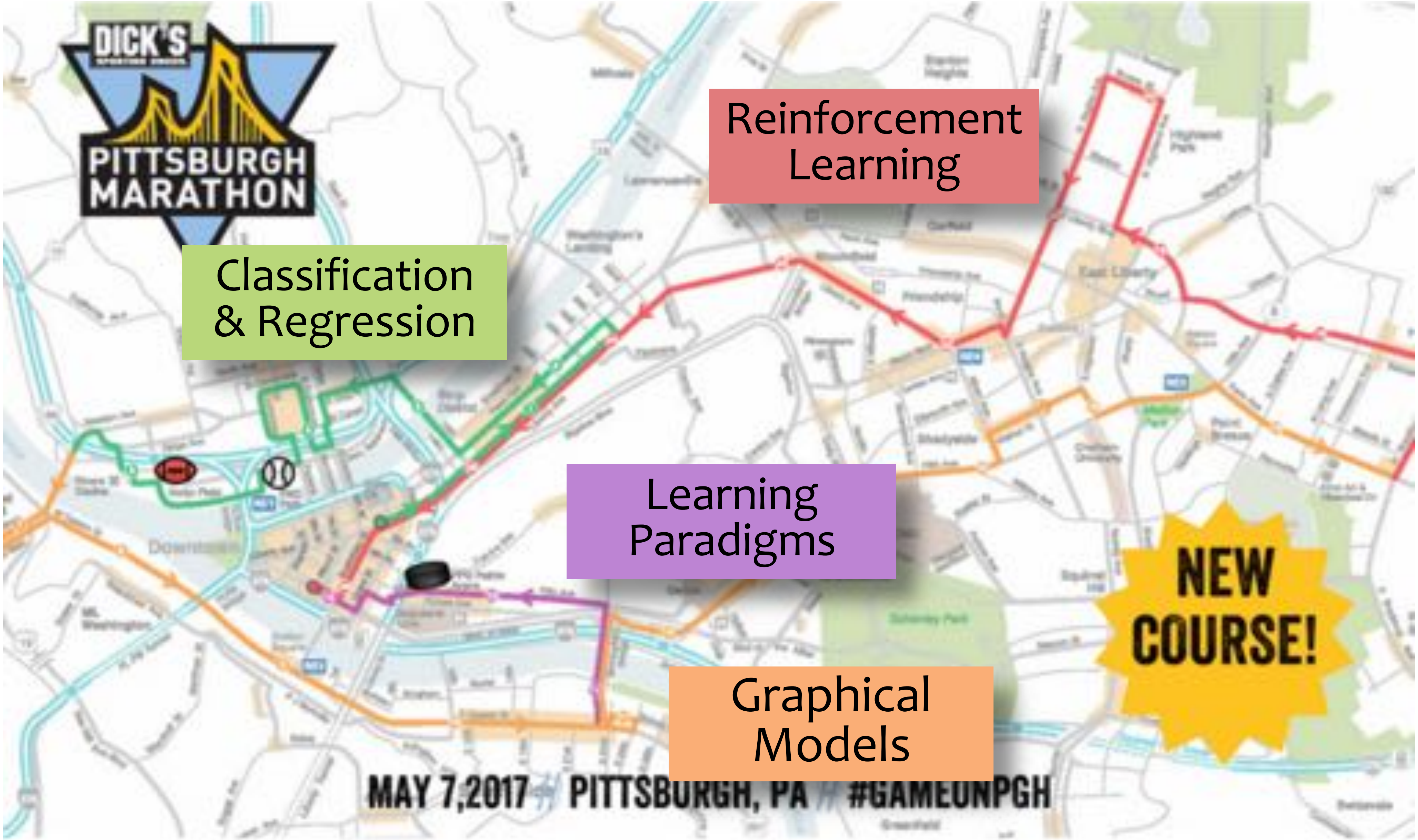
Topics for Midterm 2

- Classification
  - Binary Logistic Regression
  - Multinomial Logistic Regression
- Important Concepts
  - 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
- Unsupervised Learning
  - K-Means
  - PCA
- Other Learning Paradigms
  - SVM (large-margin)
  - Kernels
  - ~~Ensemble Methods~~
  - ~~Recommender Systems~~

DICK'S
SPORTING GOODS
PITTSBURGH
MARATHON
4268
13344



Reinforcement Learning

Classification & Regression

Learning Paradigms

Graphical Models

Material Covered **Before** Midterm Exam 2

SAMPLE QUESTIONS

Matching Game

Goal: Match the Algorithm to its Update Rule

1. SGD for Logistic Regression

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = p(y|x)$$

2. Least Mean Squares

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

3. Perceptron (next lecture)

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \text{sign}(\boldsymbol{\theta}^T \mathbf{x})$$

4.

$$\theta_k \leftarrow \theta_k + (h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)})$$

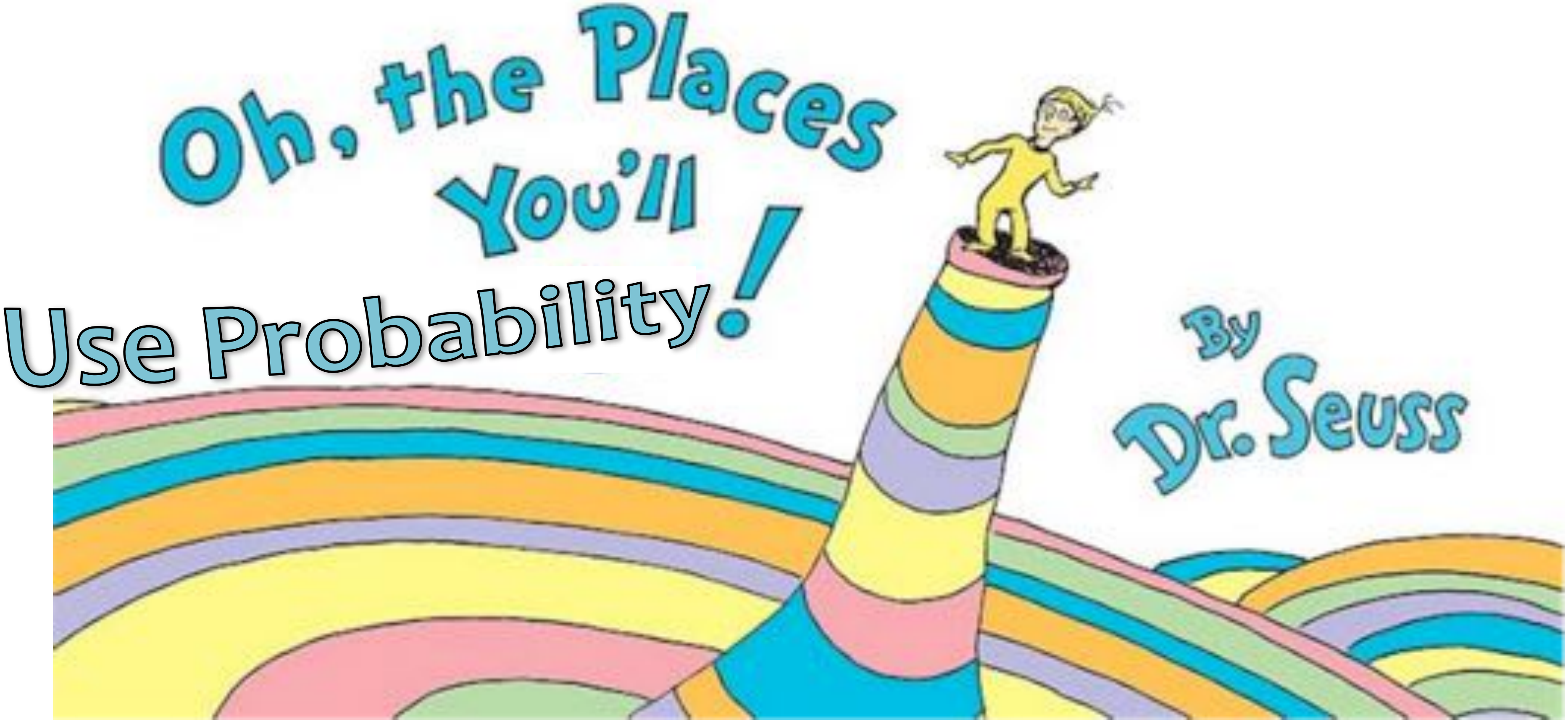
5.

$$\theta_k \leftarrow \theta_k + \frac{1}{1 + \exp \lambda(h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)})}$$

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

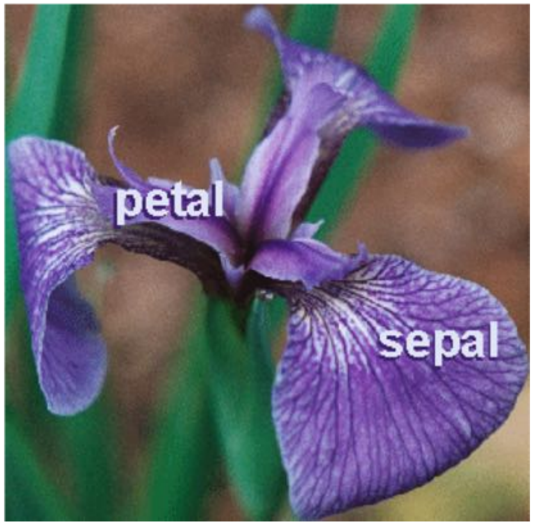
Sample Questions

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')



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.

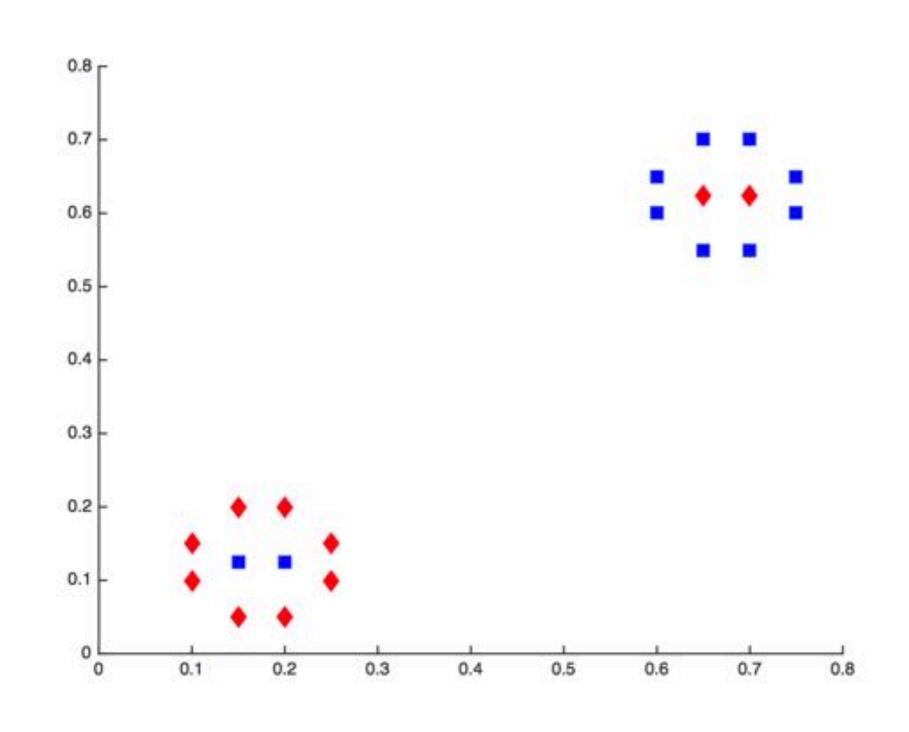


Figure 5

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?

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					

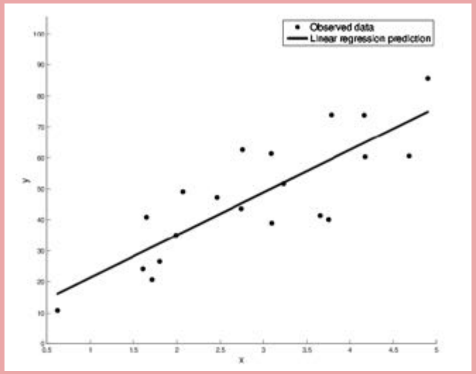
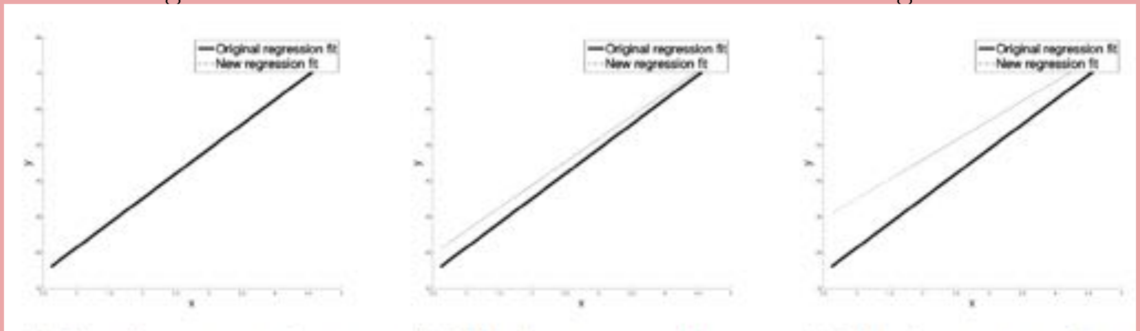


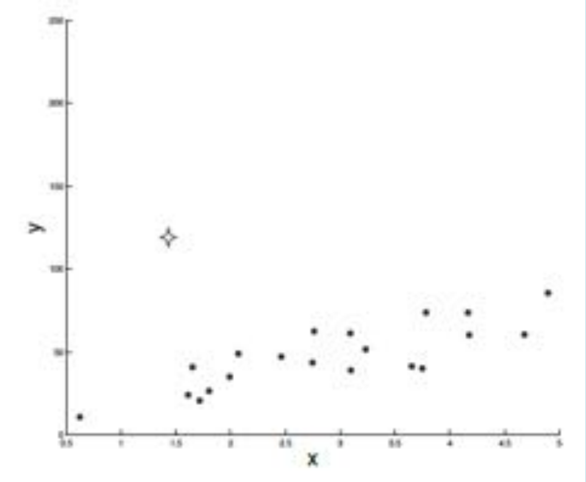
Figure 1: An observed data set and its associated regression line.



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

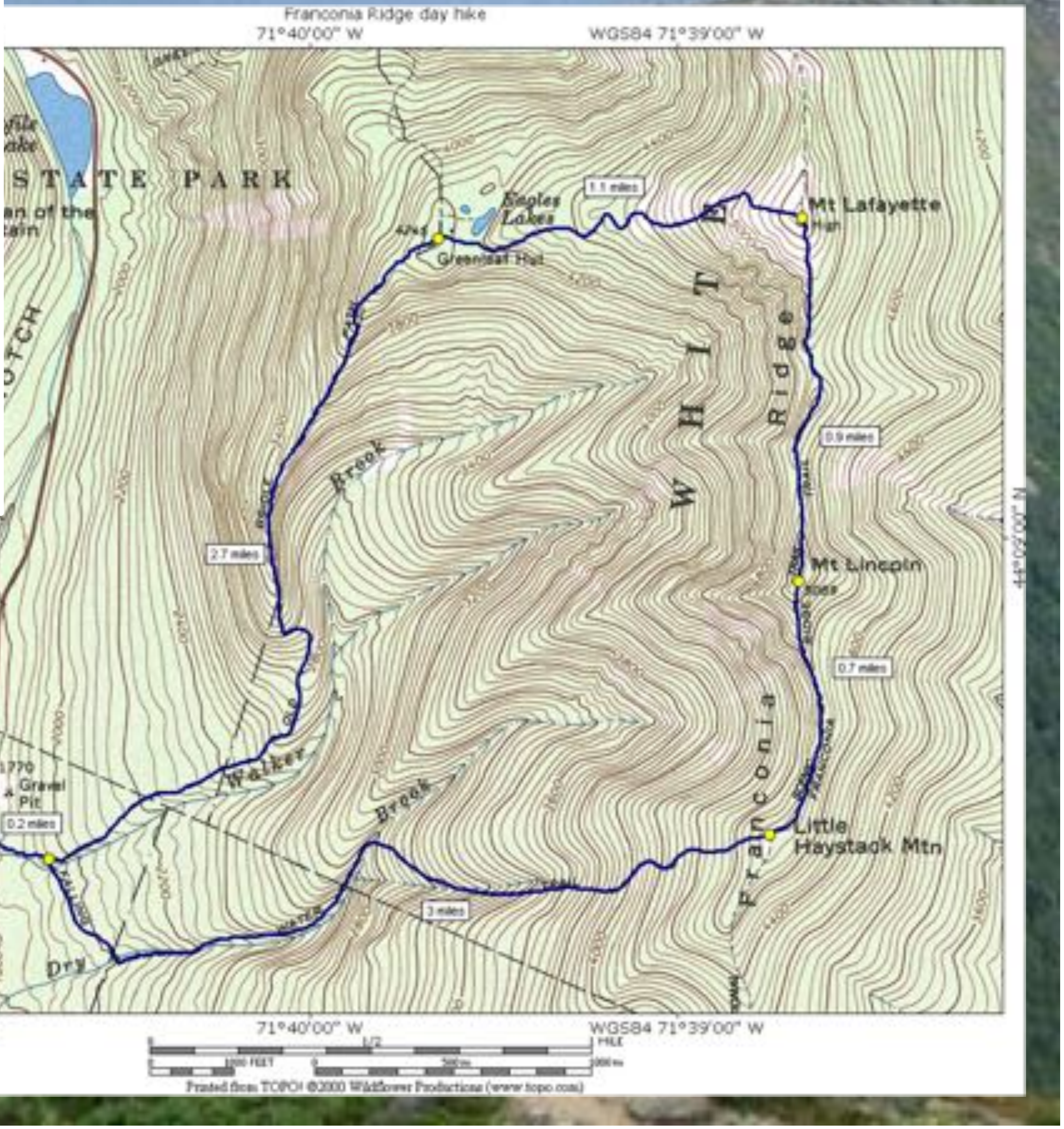
Figure 2: New regression lines for altered data sets S^{new}.

Dataset



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

Topographical Maps



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					

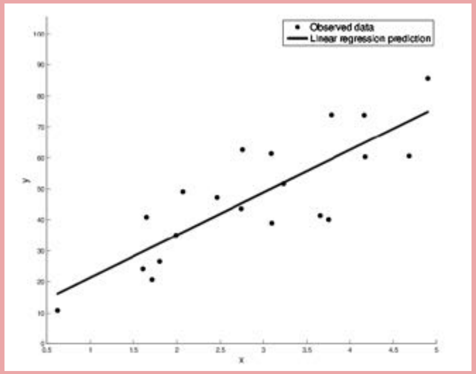
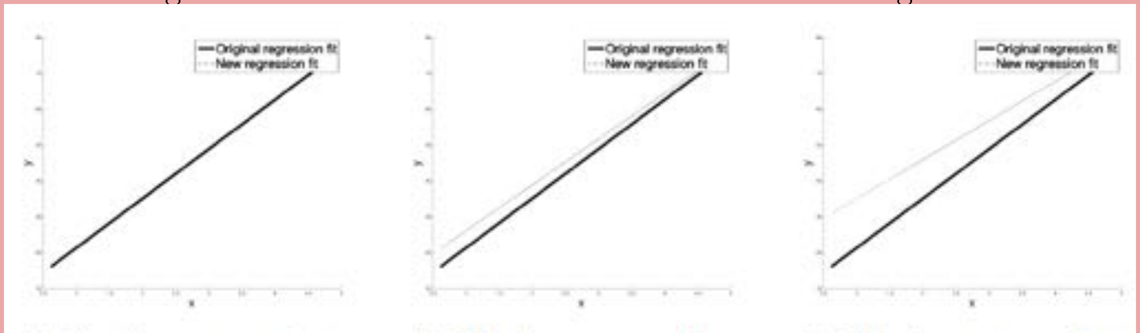


Figure 1: An observed data set and its associated regression line.



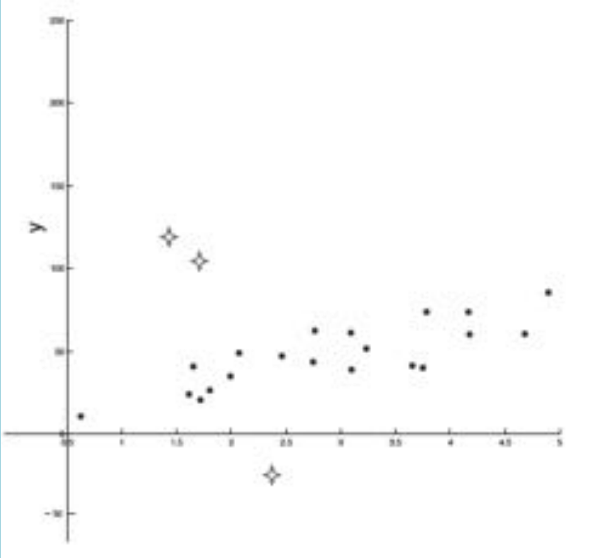
(a) Old and new regression lines.

(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					

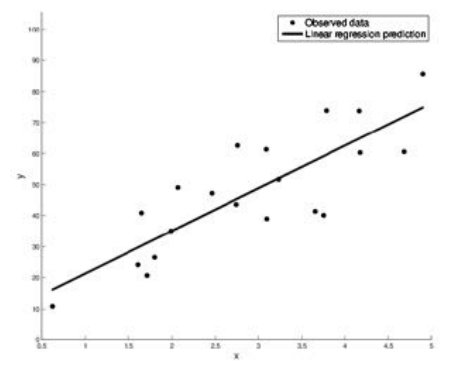
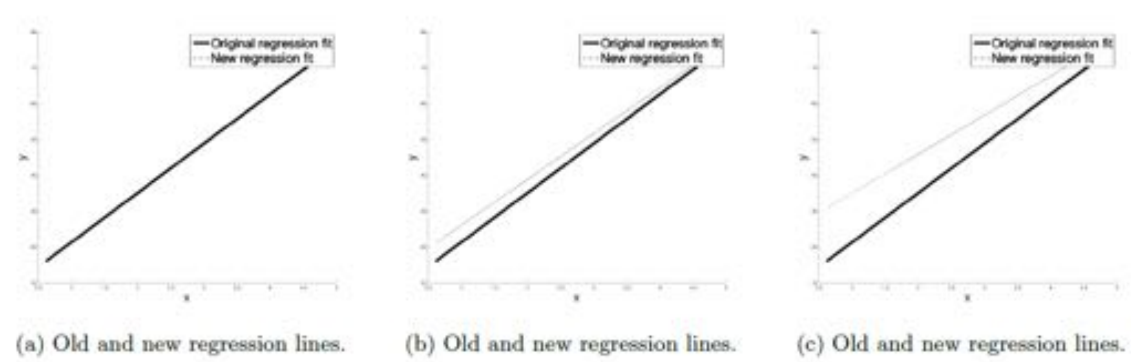


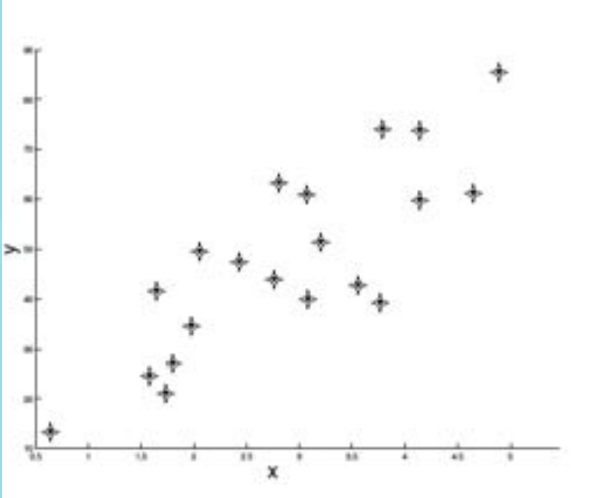
Figure 1: An observed data set and its associated regression line.



(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					

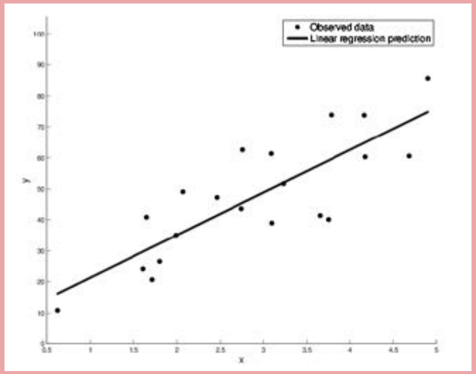
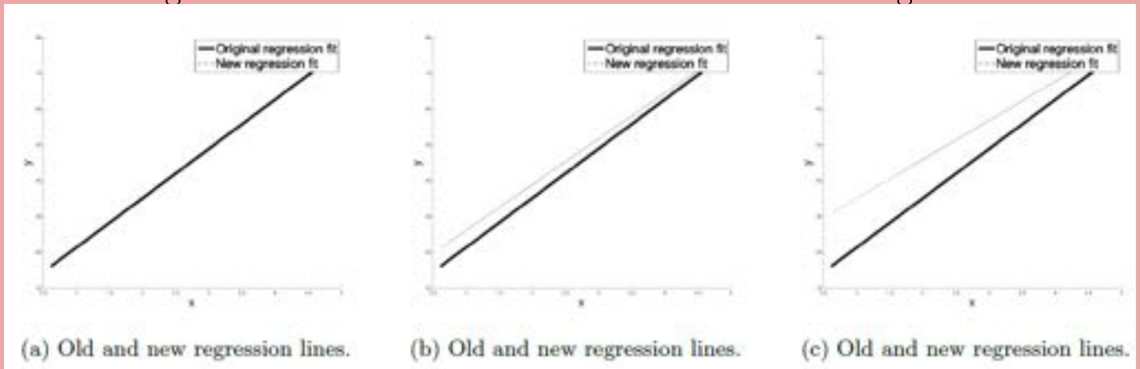


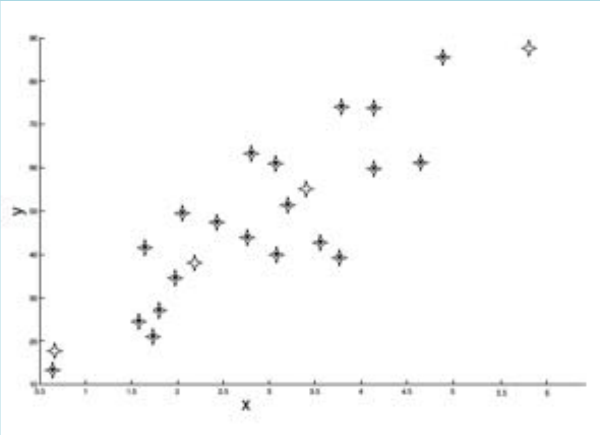
Figure 1: An observed data set and its associated regression line.



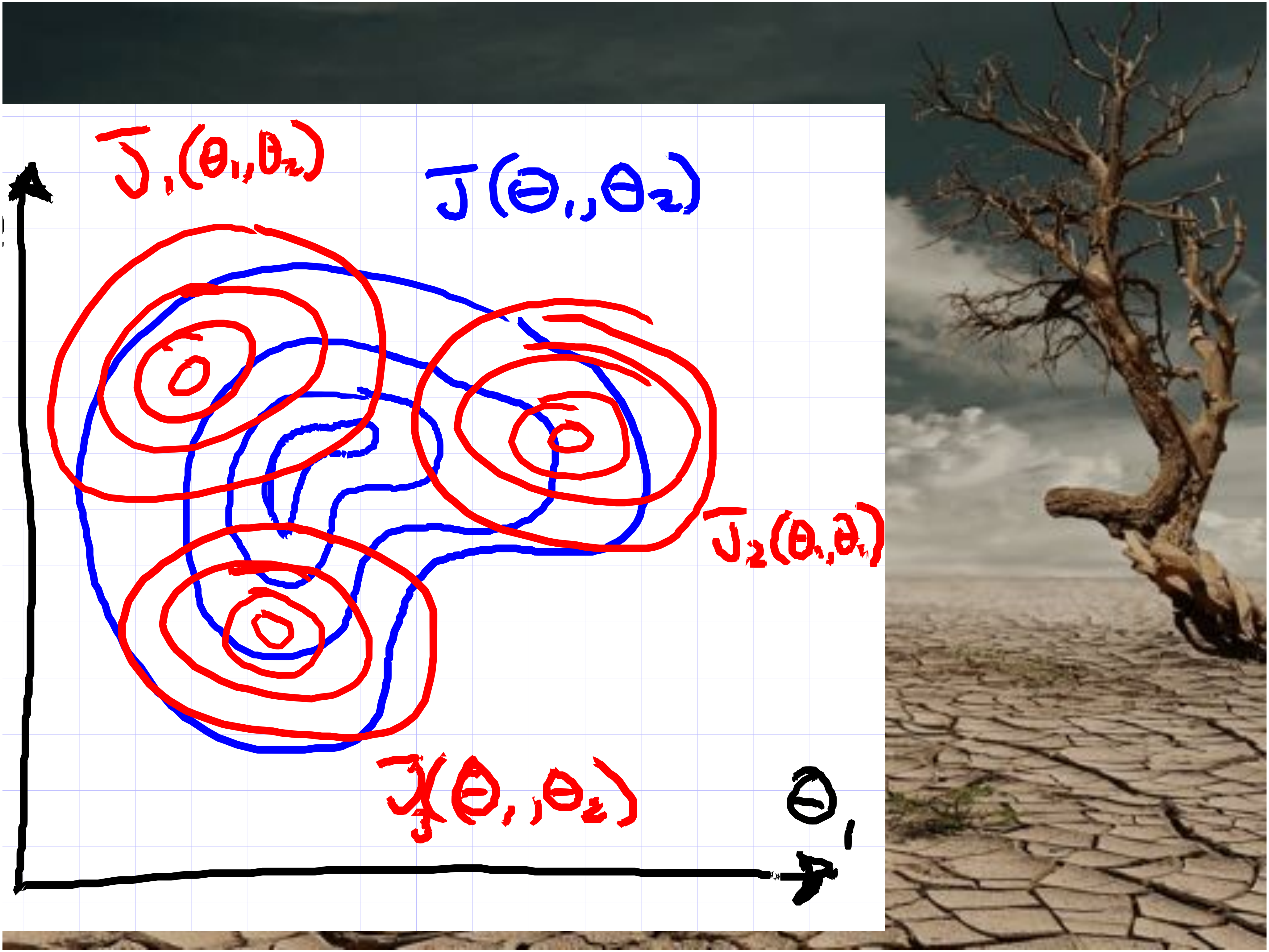
(a) Old and new regression lines. (b) Old and new regression lines. (c) Old and new regression lines.

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.


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

Robotic Farming

	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

polar bears

sea lions

sharks

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

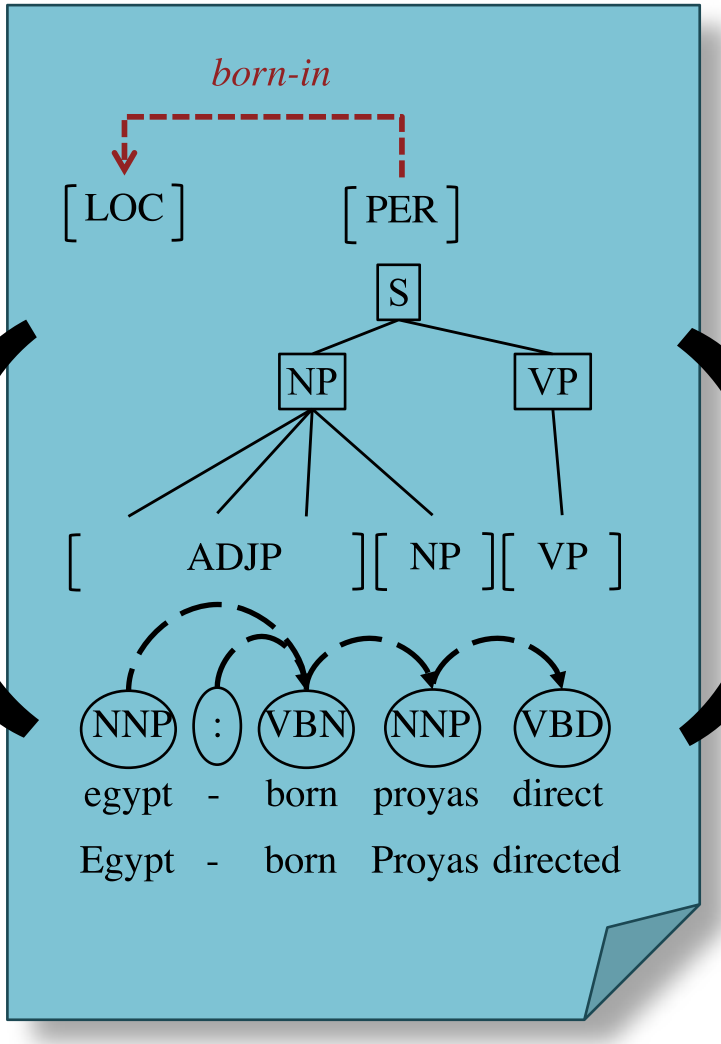
$$\nabla \ell(w) = \sum_{i=1}^{n} (y_i - p(y_i|x_i; w)) x_i.$$

(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 $\hat{w}$?

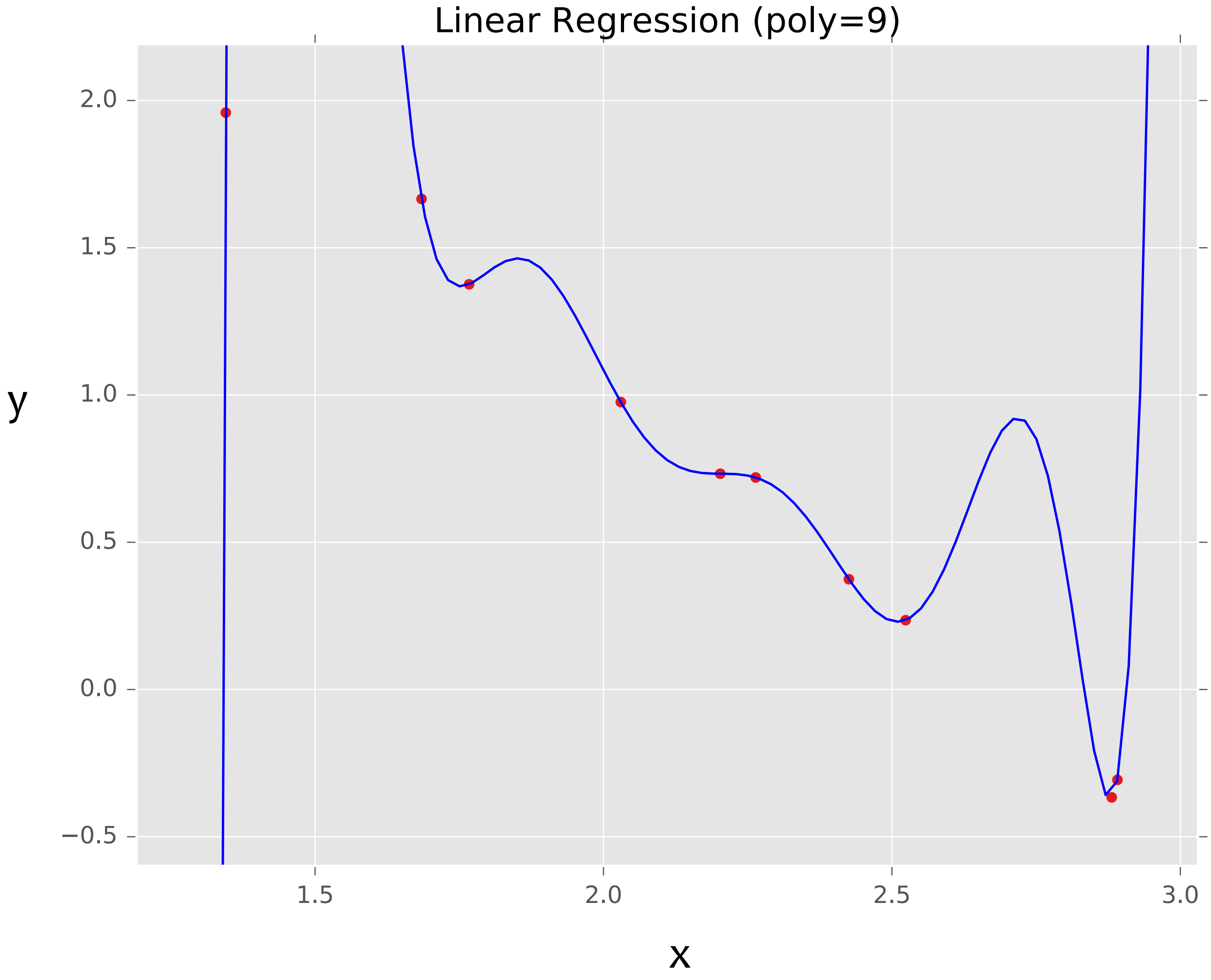
Handcrafted Features

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



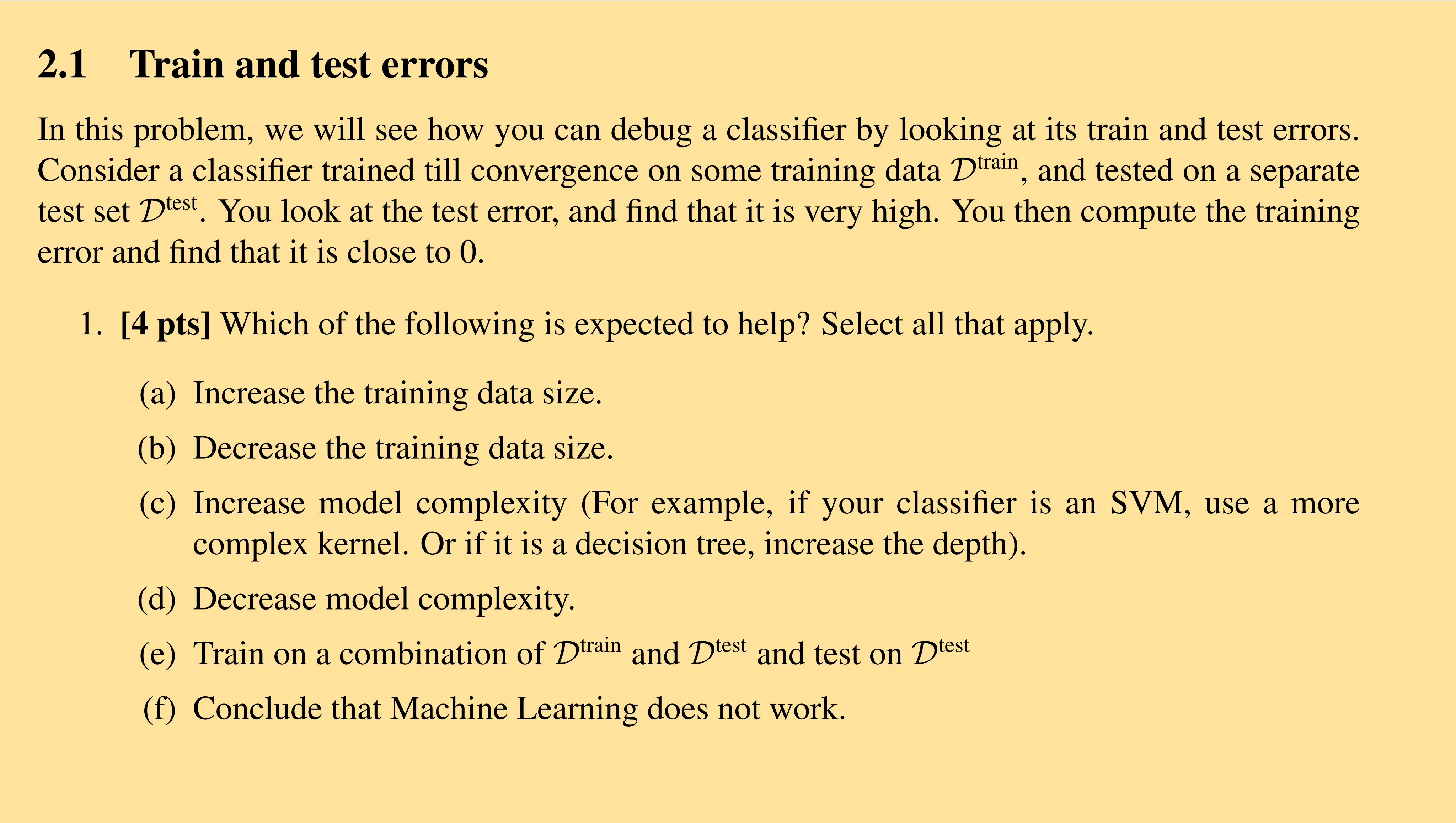
Example: Linear Regression

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

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 $\mathcal{D}^{\text{test}}$

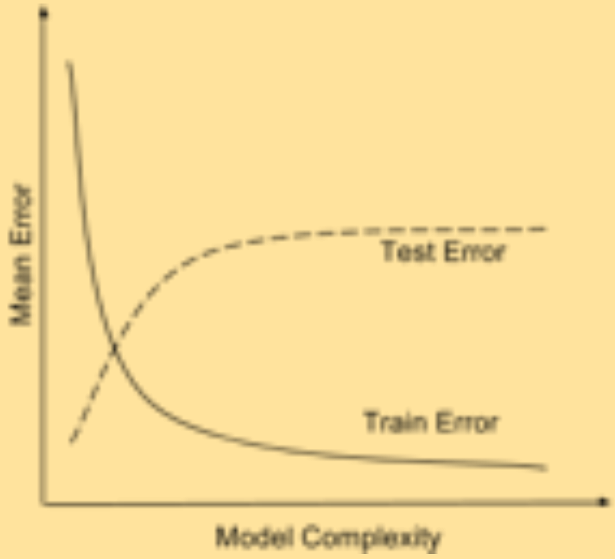
   (f) Conclude that Machine Learning does not work.

Samples Questions

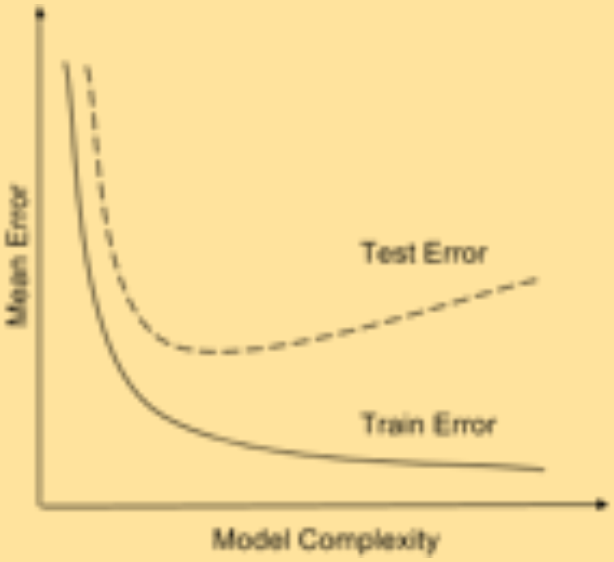
2.1 Train and test errors

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**.

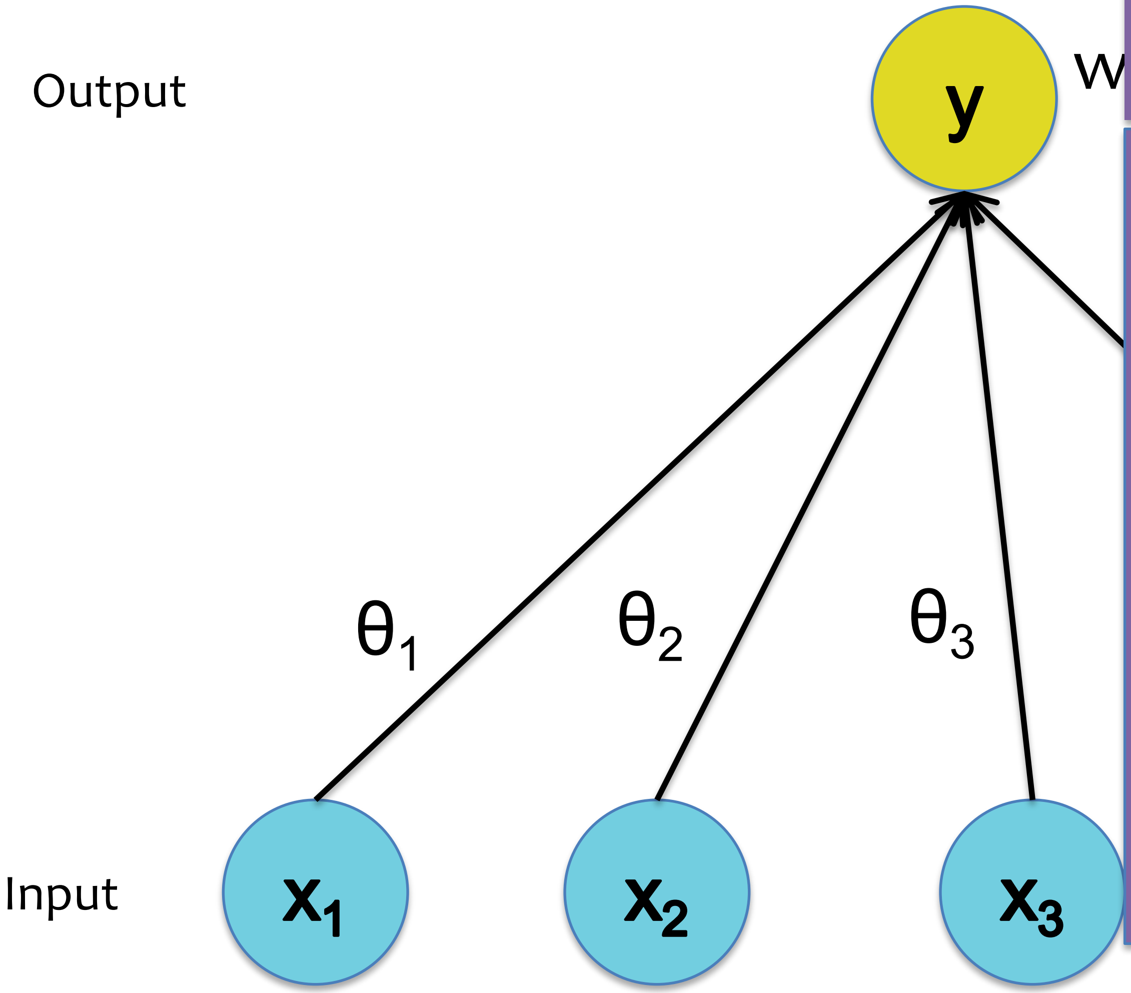
(a) [2 pts.] Consider two datasets $D^{(1)}$ and $D^{(2)}$ where $D^{(1)} = \{(x_1^{(1)}, y_1^{(1)}), ..., (x_n^{(1)}, y_n^{(1)})\}$ and $D^{(2)} = \{(x_1^{(2)}, y_1^{(2)}), ..., (x_m^{(2)}, y_m^{(2)})\}$ such that $x_i^{(1)} \in \mathbb{R}^{d_1}$, $x_i^{(2)} \in \mathbb{R}^{d_2}$. Suppose $d_1 > d_2$ 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

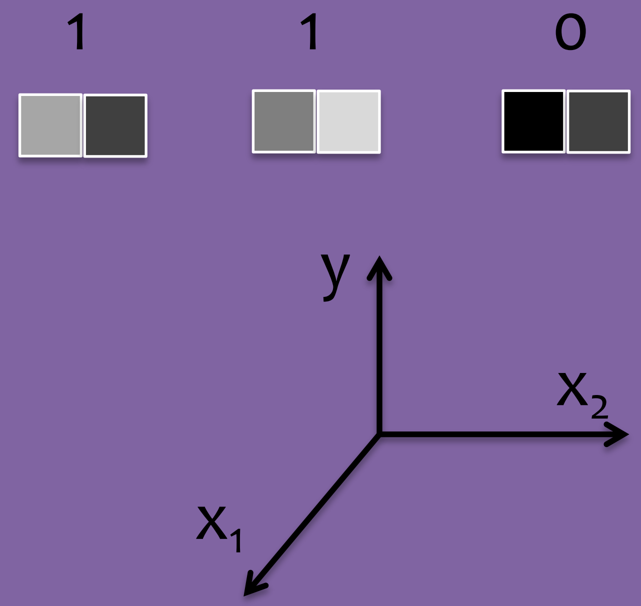
Logistic Regression

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

w



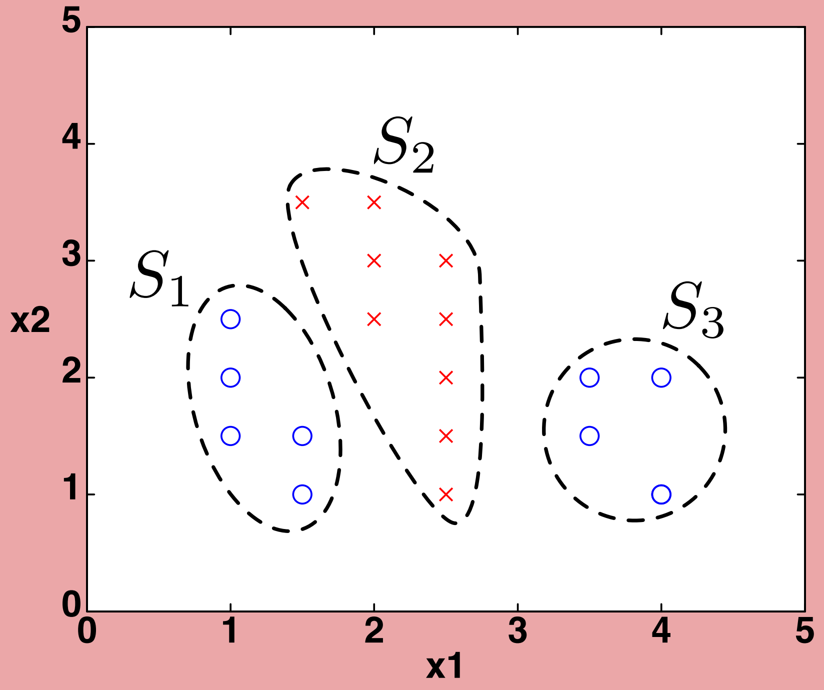
In-Class Example



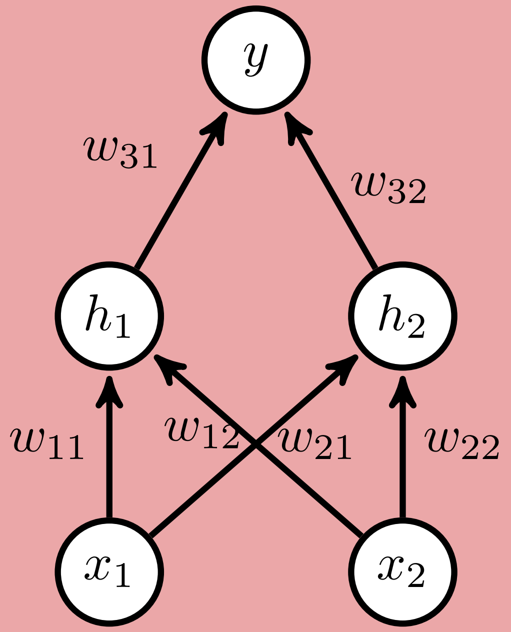
Sample Questions

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.

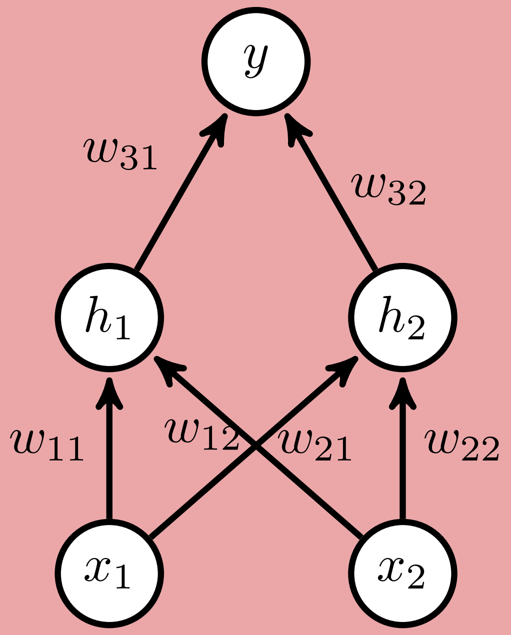


(b) The neural network architecture

Sample Questions

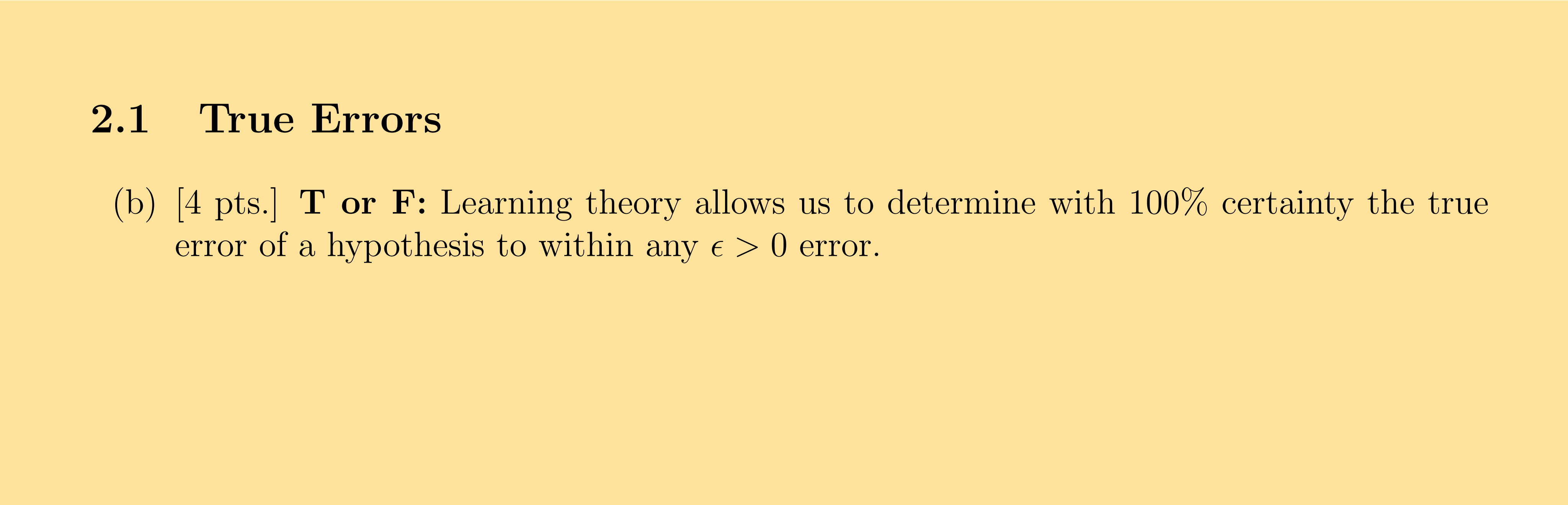
Neural Networks

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

Samples Questions

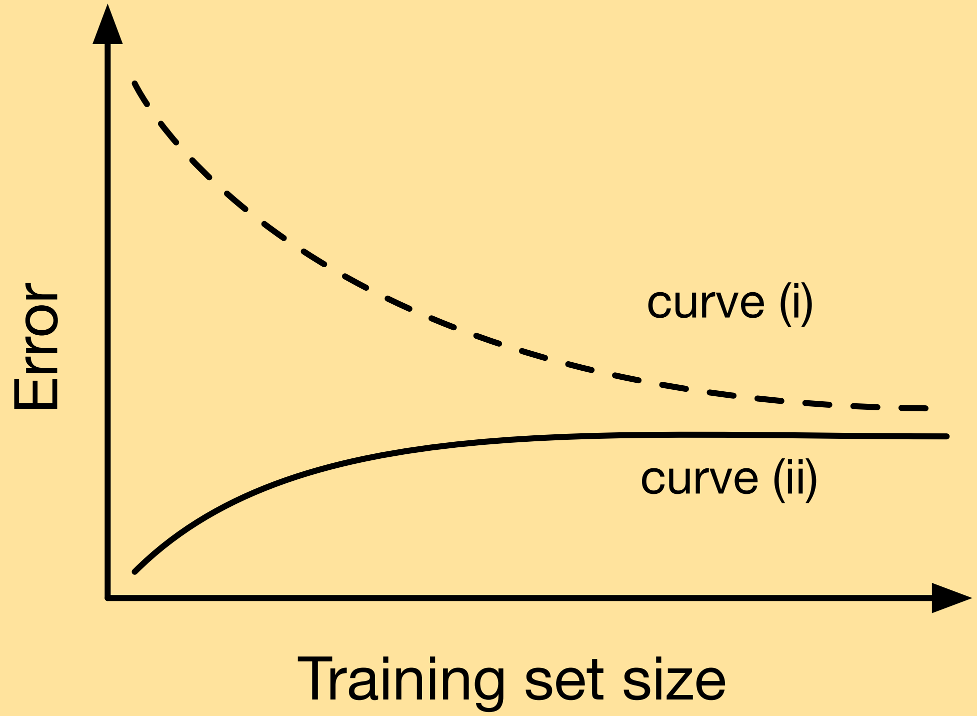


2.1 True Errors

(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

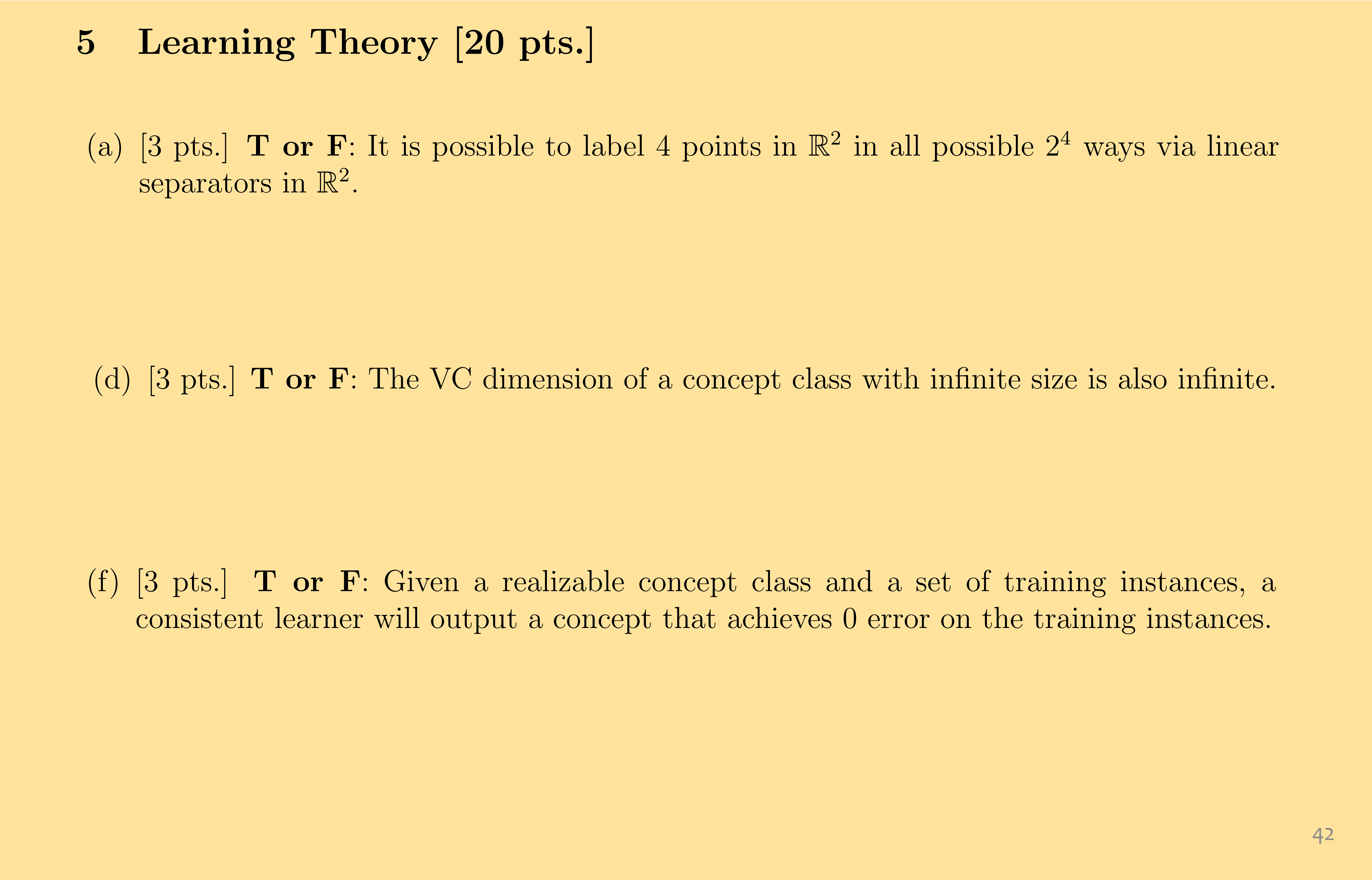
2.2 Training Sample Size



(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



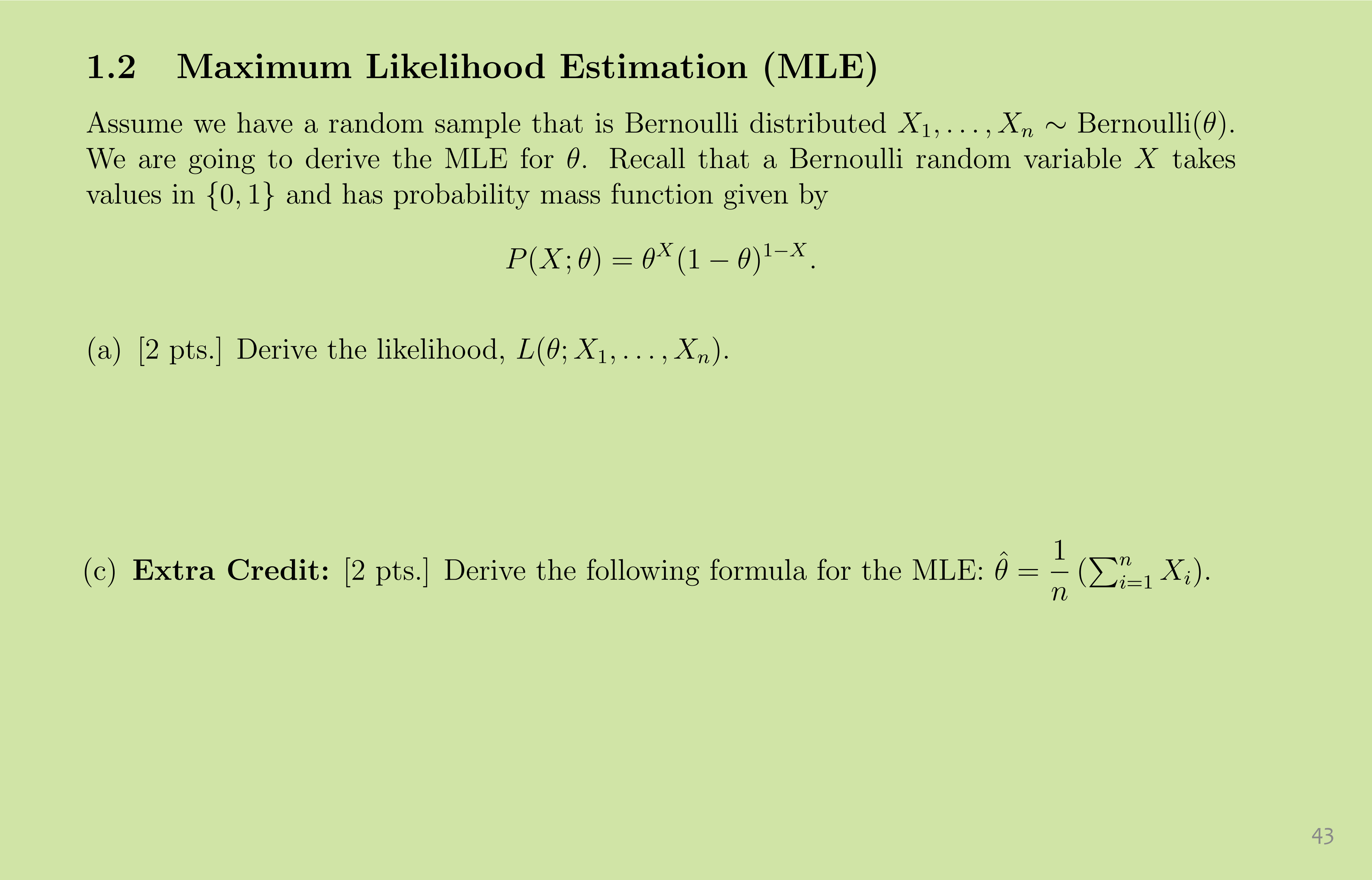
5 Learning Theory [20 pts.]

(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.

Sample Questions



1.2 Maximum Likelihood Estimation (MLE)

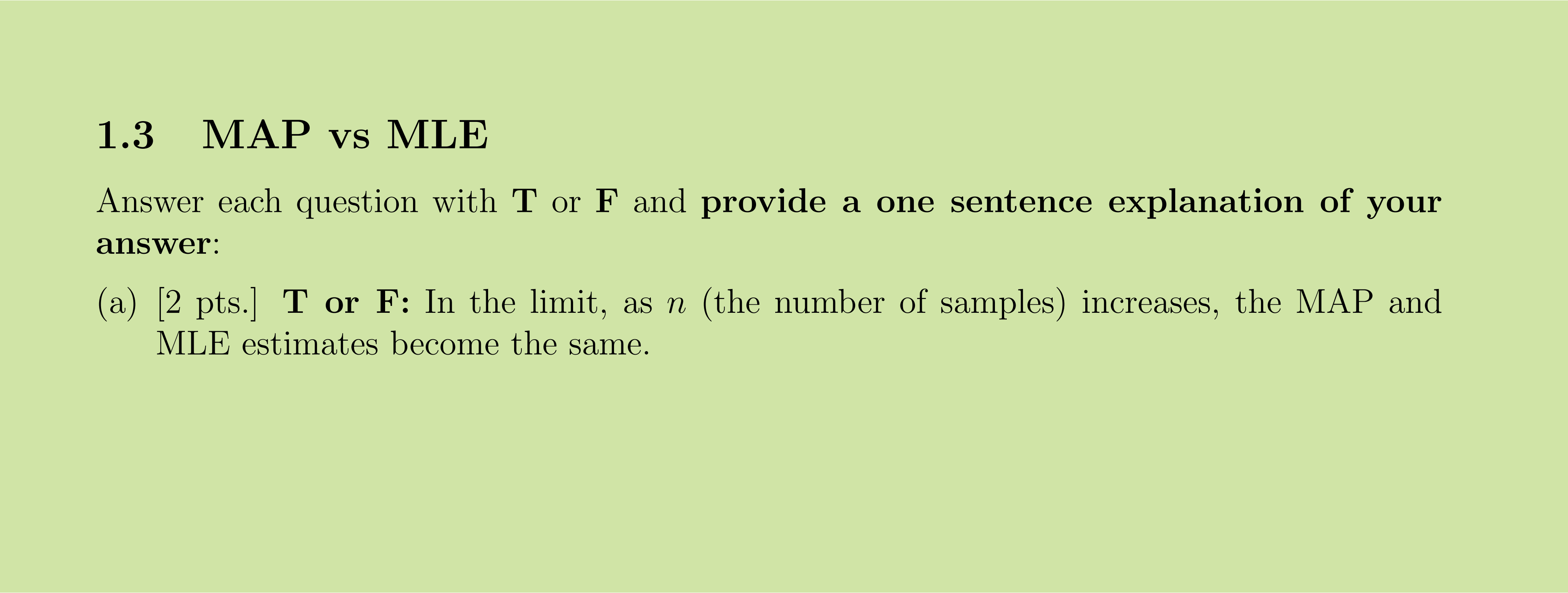
Assume we have a random sample that is Bernoulli distributed $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$. We are going to derive the MLE for θ. Recall that a Bernoulli random variable X takes values in $\{0, 1\}$ and has probability mass function given by

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

(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: $\hat{\theta} = \frac{1}{n} \left(\sum_{i=1}^{n} X_i\right)$.

Sample Questions

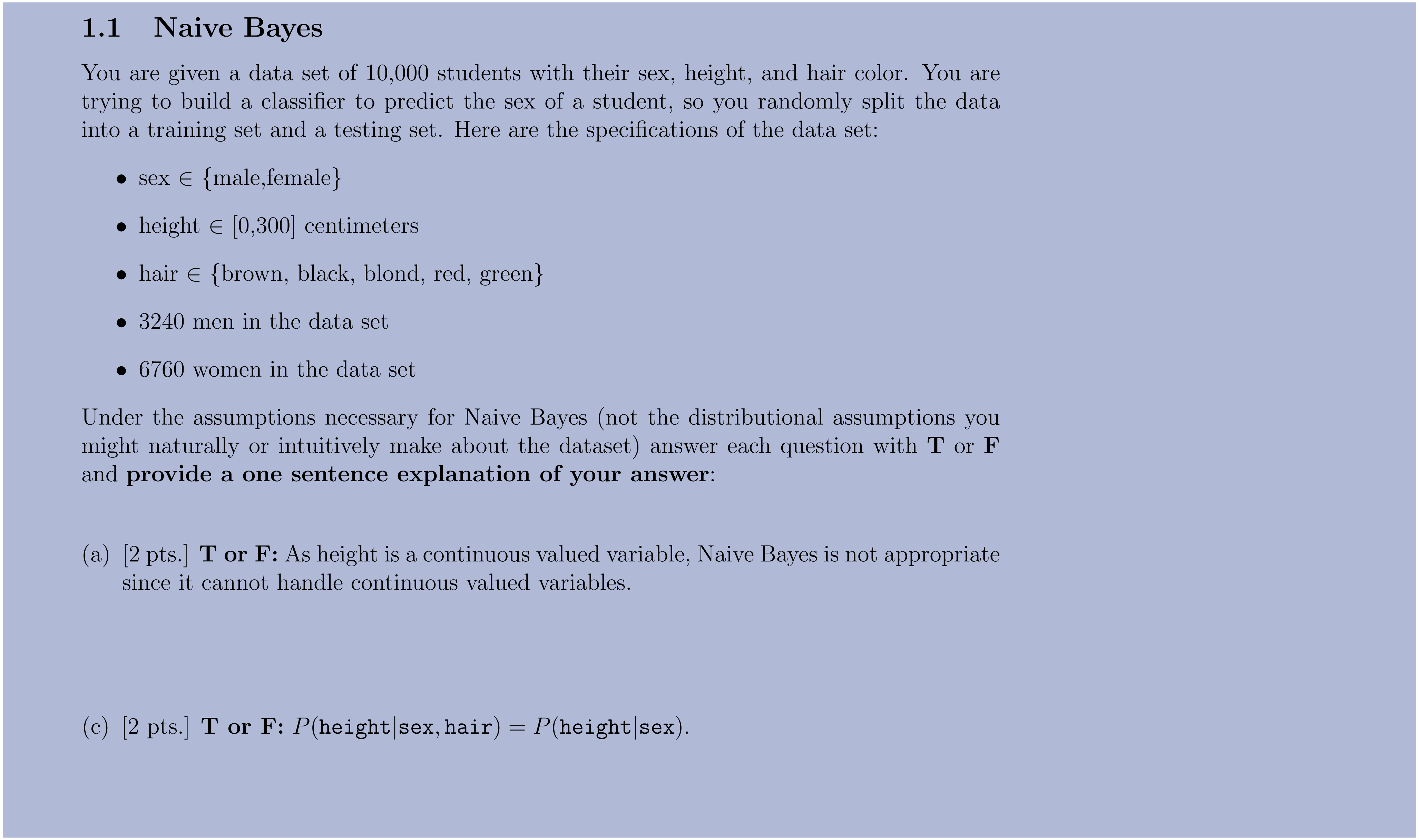


1.3 MAP vs MLE

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.

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**:

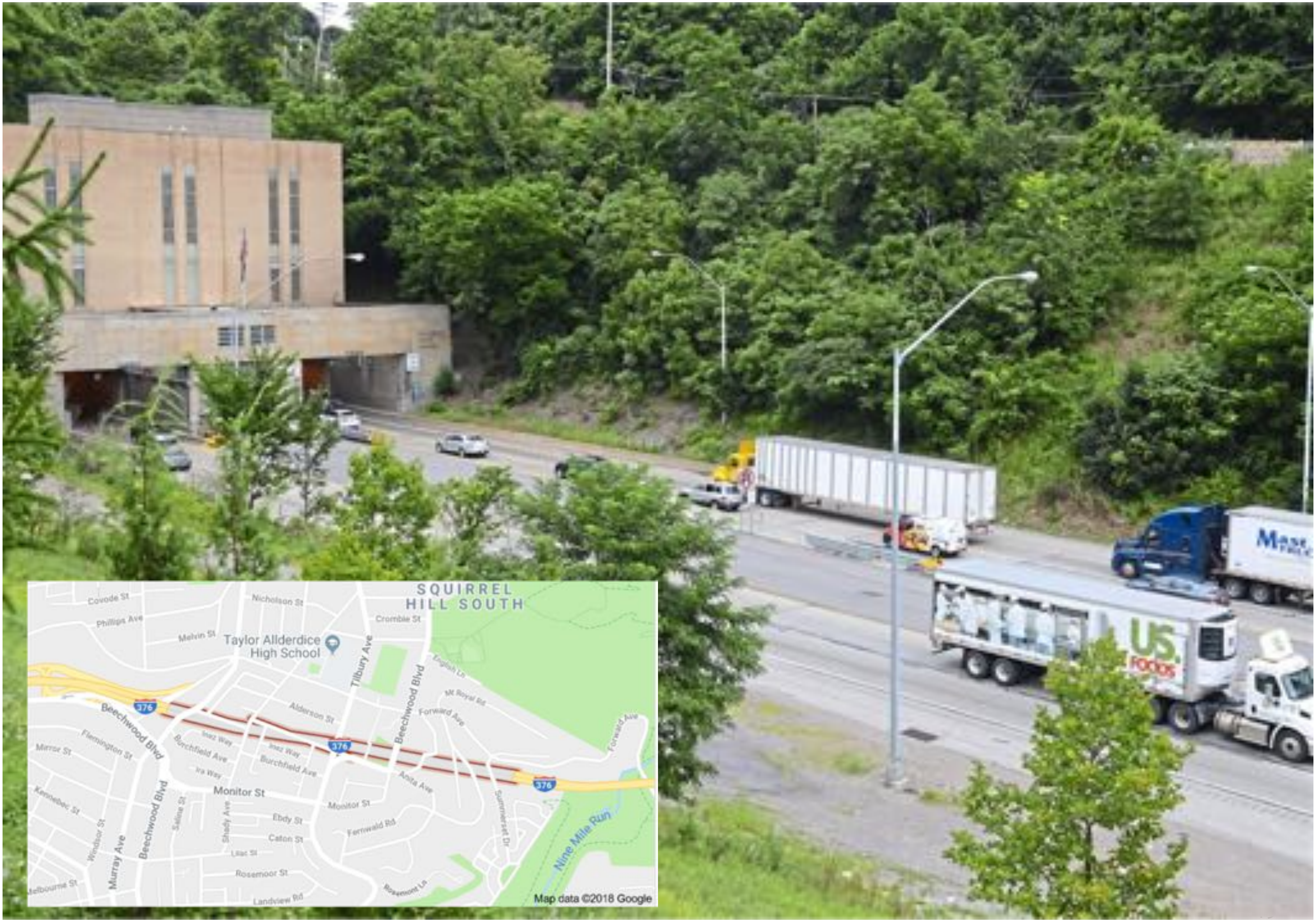
(a) [2 pts.] **T or F:** As height is a continuous valued variable, Naive Bayes is not appropriate since it cannot handle continuous valued variables.

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

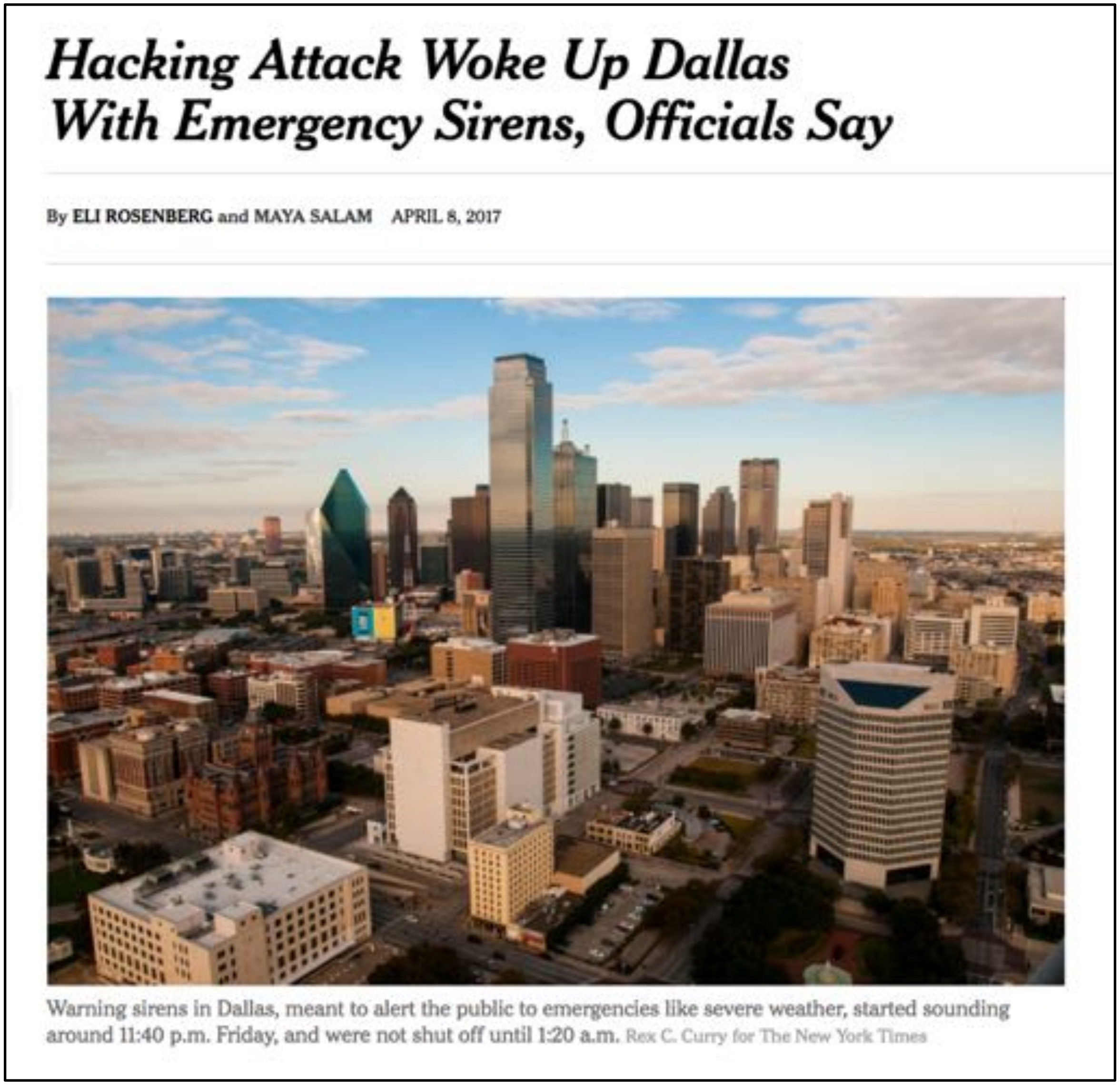
Material Covered **After** Midterm Exam 2

SAMPLE QUESTIONS

Totoro’s Tunnel


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

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.

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., $R, S, E, A \in \{0, 1\}$.

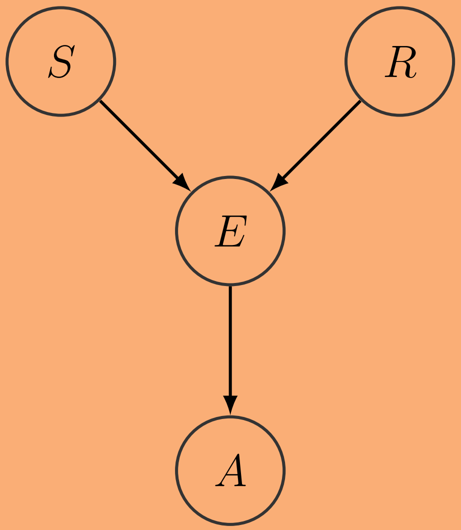


Figure 5: Directed graphical model for problem 5.

Sample Questions

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

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., $R, S, E, A \in \{0, 1\}$.

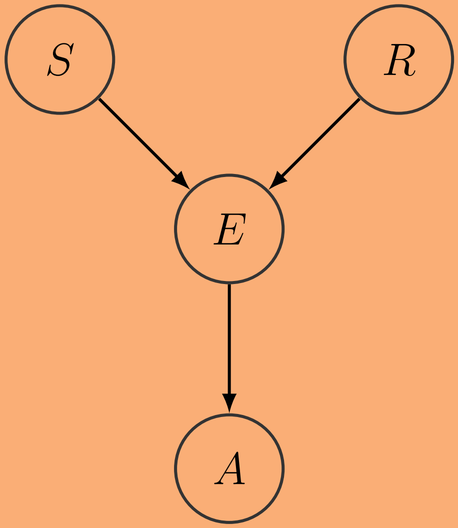


Figure 5: Directed graphical model for problem 5.

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., $R, S, E, A \in \{0, 1\}$.

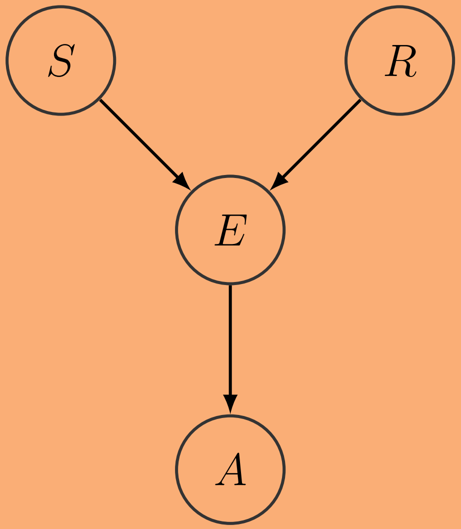
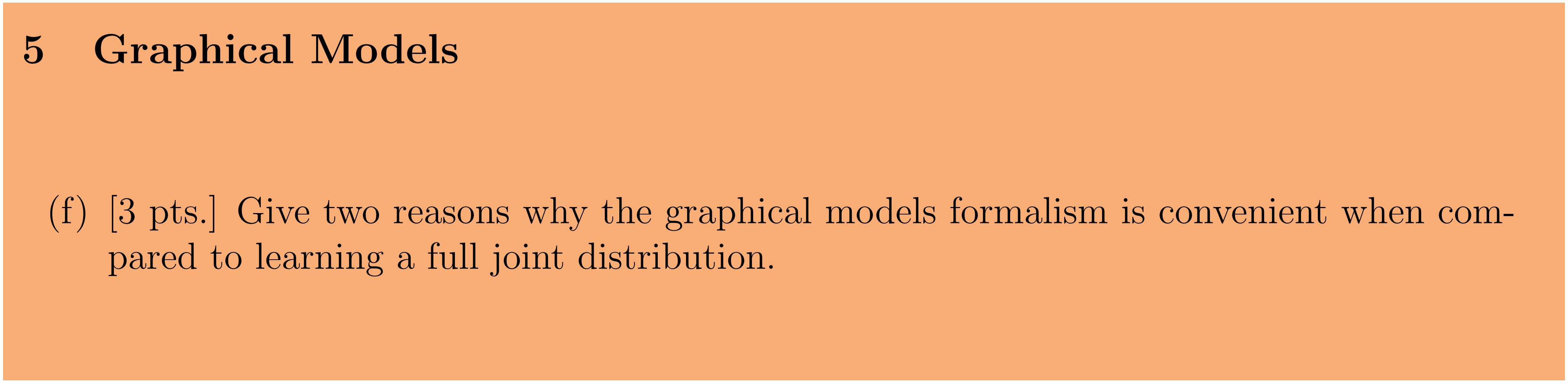


Figure 5: Directed graphical model for problem 5.

Sample Questions

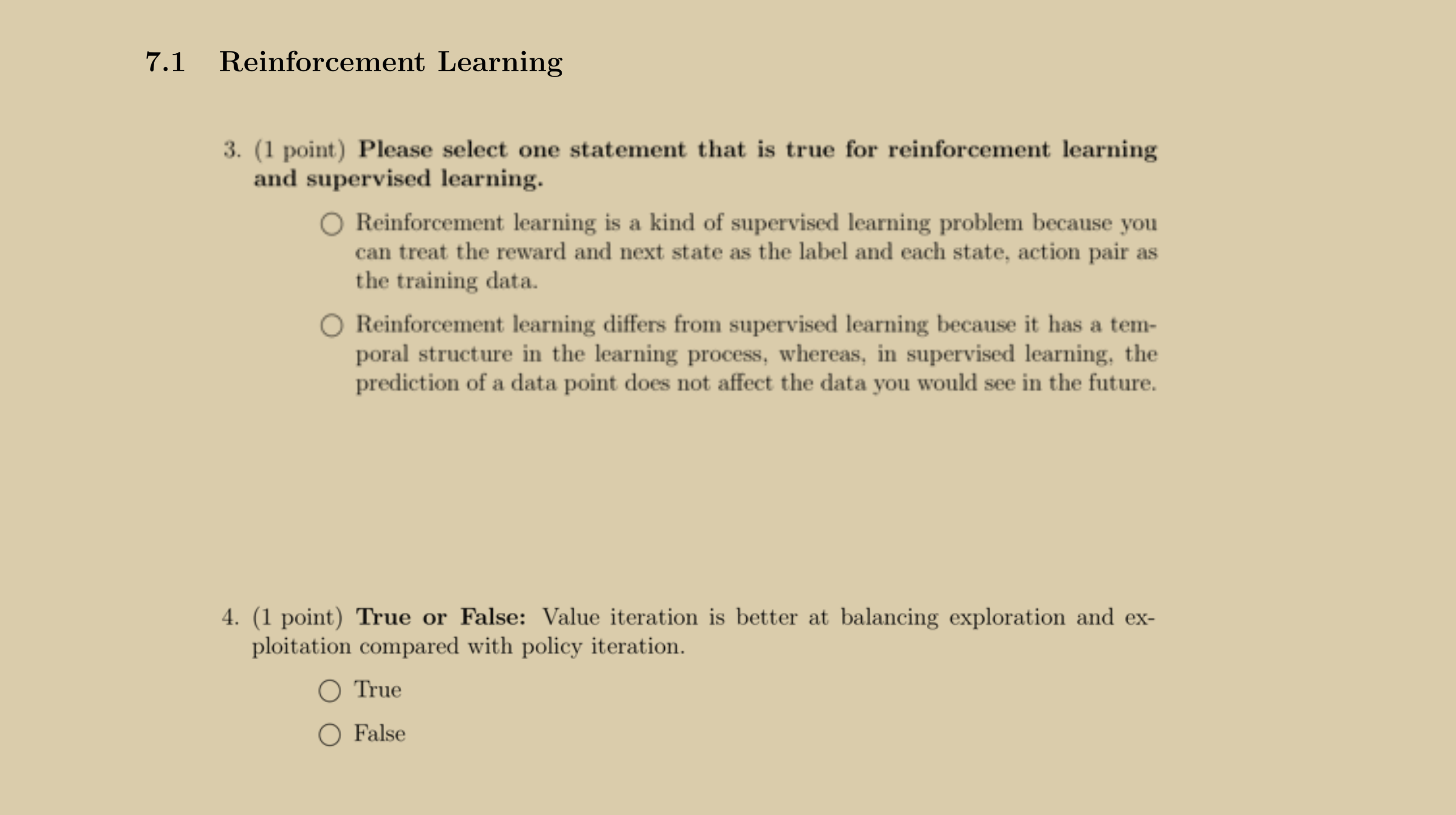


5 Graphical Models

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

Example: Path Planning

Sample Questions



7.1 Reinforcement Learning

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

   ○ 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.

   ○ 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

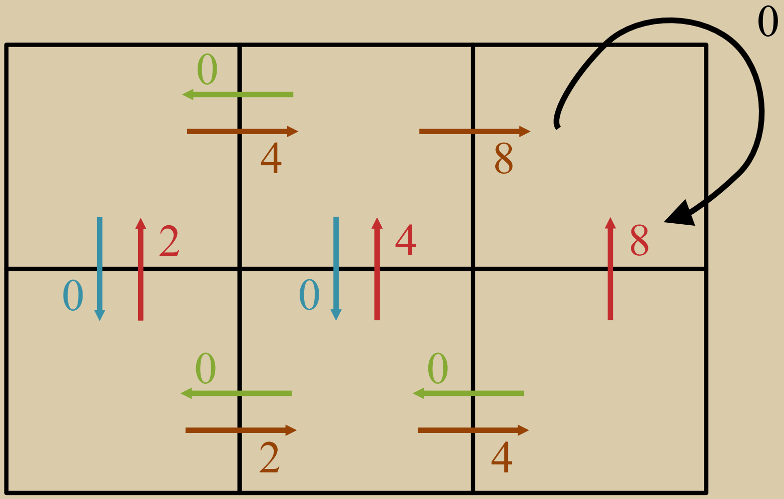
Sample Questions

7.1 Reinforcement Learning

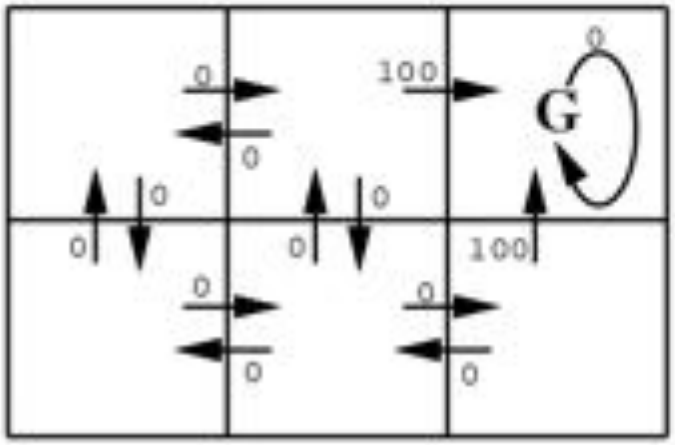
1. For the R(s,a) values shown on the arrows below, what is the corresponding optimal policy? Assume the discount factor is 0.1

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

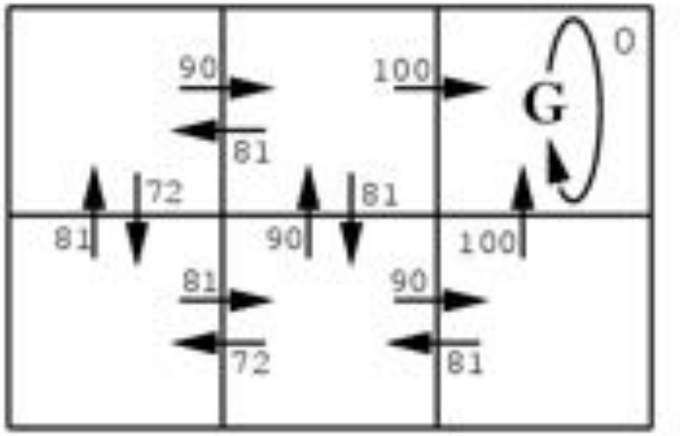
3. For the R(s,a) values shown on the arrows below, which are the corresponding Q*(s,a) values? Assume the discount factor is 0.1



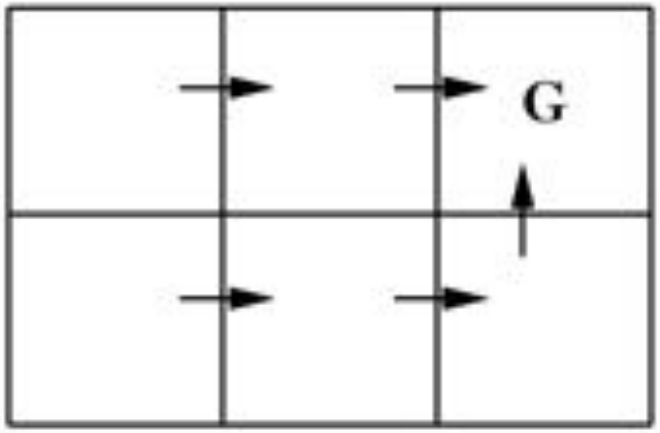
Example: Robot Localization



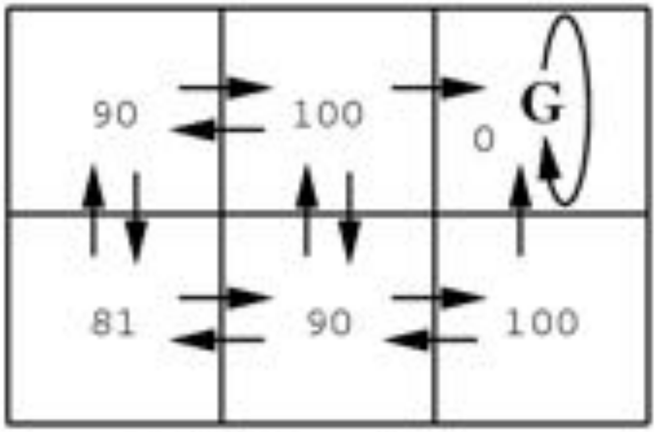
$r(s, a)$ (immediate reward) values



$Q(s, a)$ values



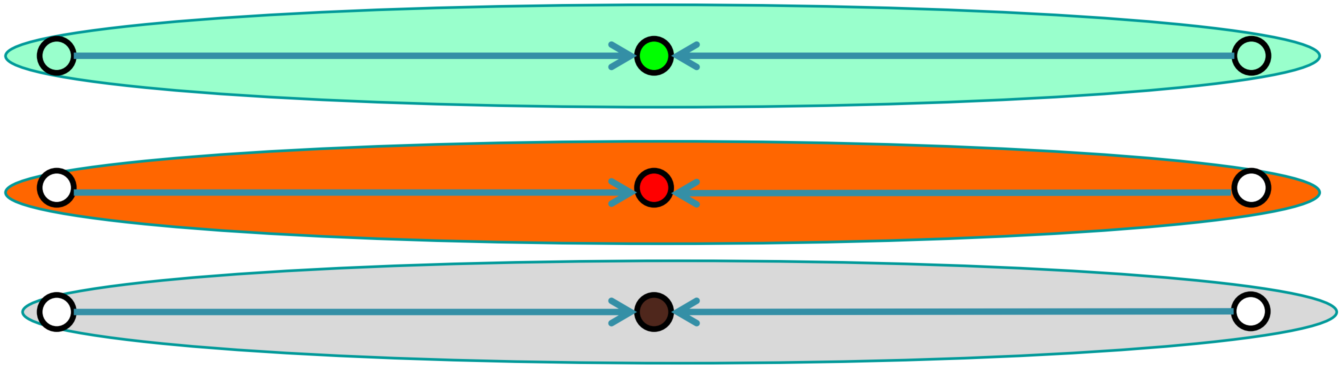
One optimal policy



$V^*(s)$ values

Figure from Tom Mitchell

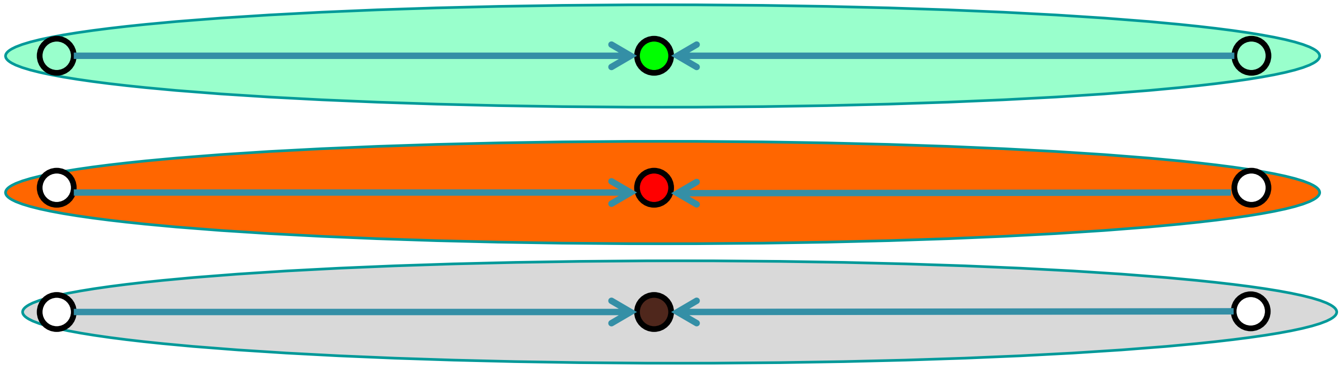
Lloyd’s method: Performance



It always converges, but it may converge at a local optimum that is different from the global optimum, and in fact could be arbitrarily worse in terms of its score.

Slide courtesy of Nina Balcan

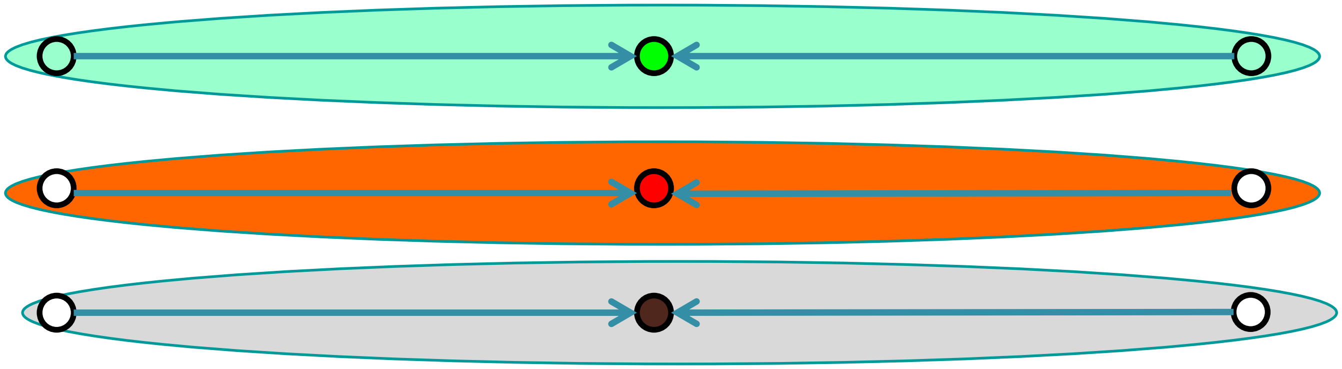
Lloyd’s method: Performance



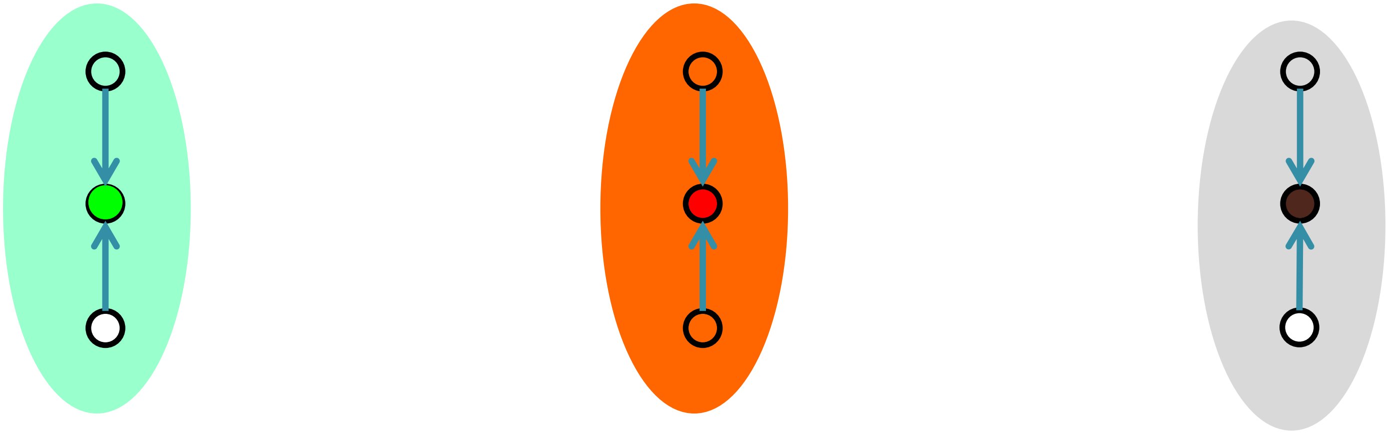
Local optimum: every point is assigned to its nearest center and every center is the mean value of its points.

Slide courtesy of Nina Balcan

Lloyd's method: Performance

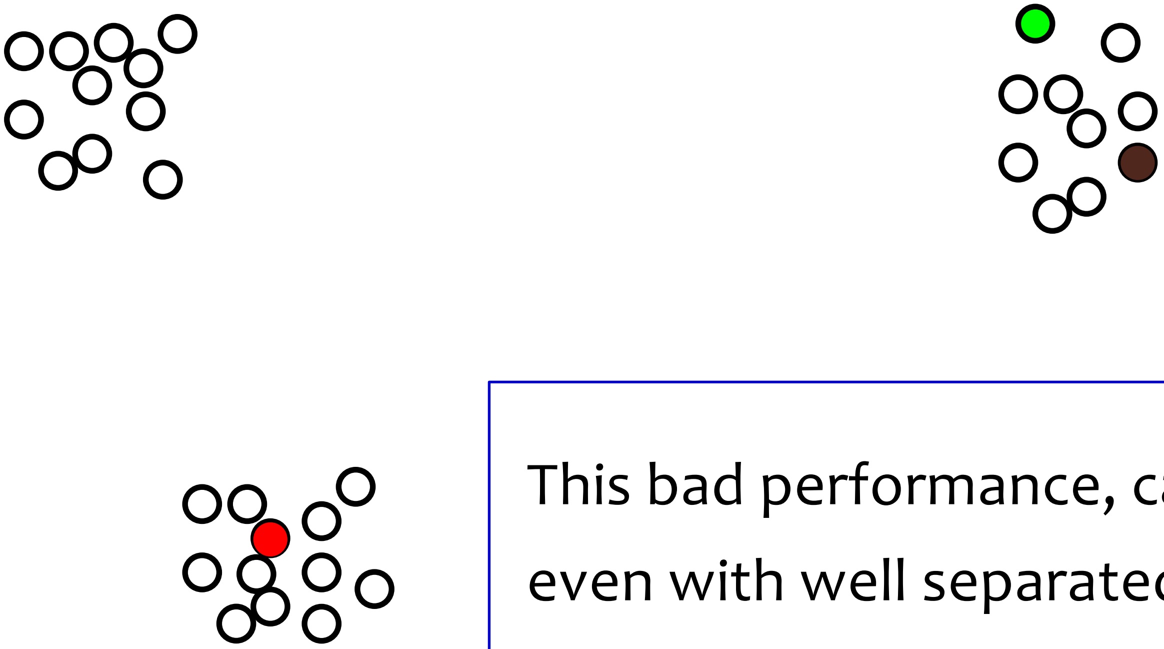


.It is arbitrarily worse than optimum solution….



Slide courtesy of Nina Balcan

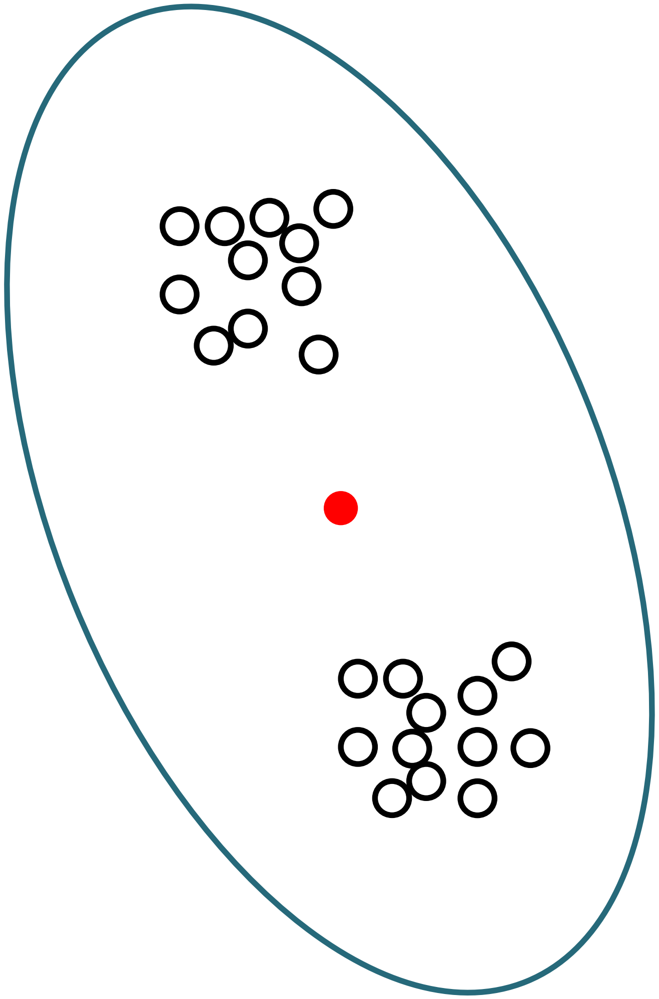
Lloyd’s method: Performance

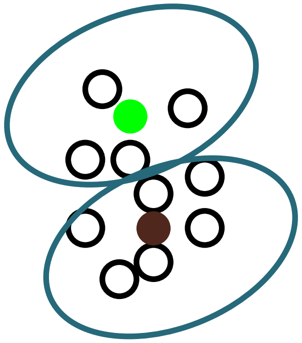


This bad performance, can happen even with well separated Gaussian clusters.

Slide courtesy of Nina Balcan

Lloyd’s method: Performance



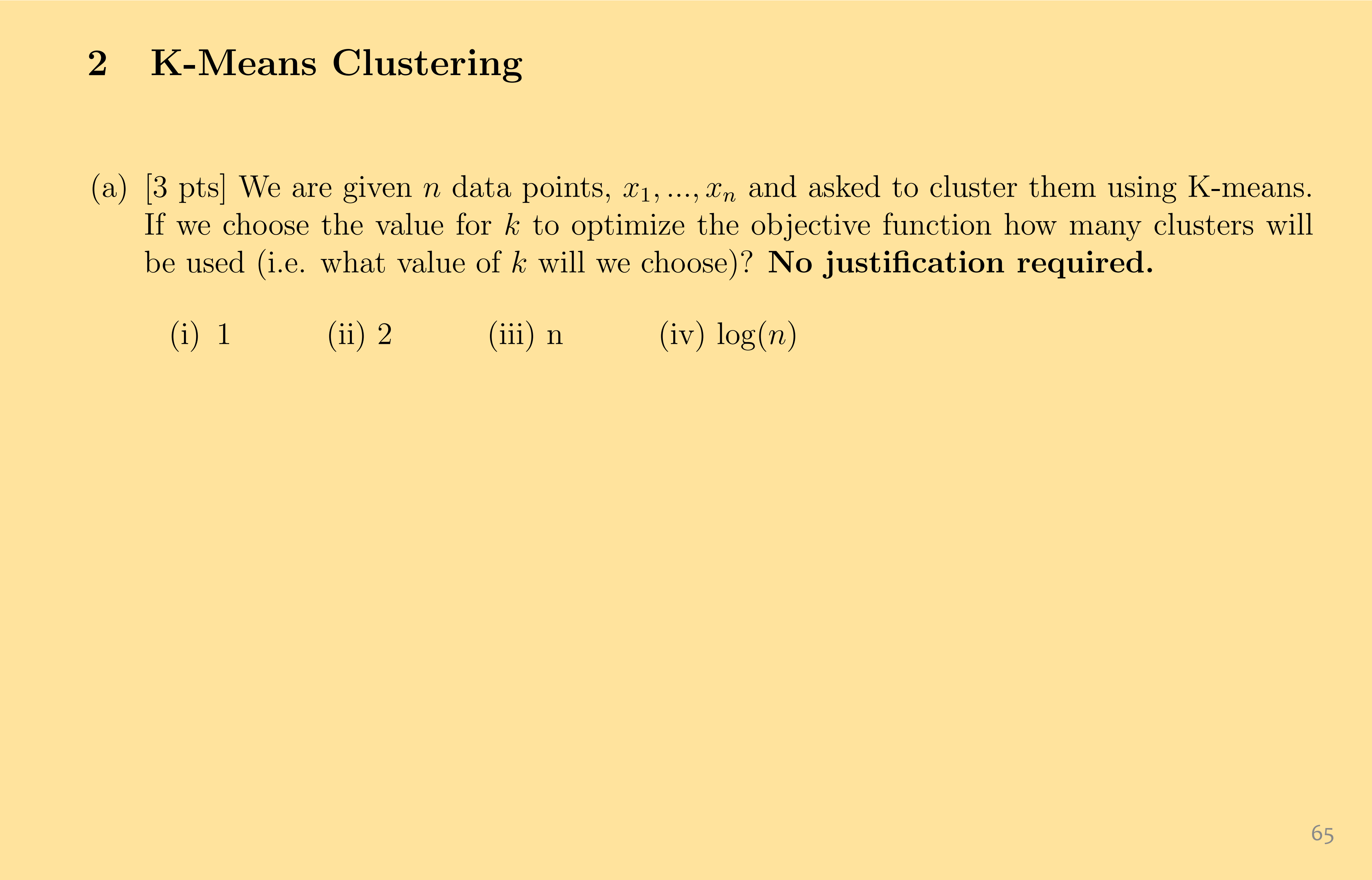


This bad performance, can happen even with well separated Gaussian clusters.

Some Gaussian are combined……

Slide courtesy of Nina Balcan

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 (ii) 2 (iii) n (iv) $\log(n)$

Samples Questions

2.2 Lloyd's algorithm

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

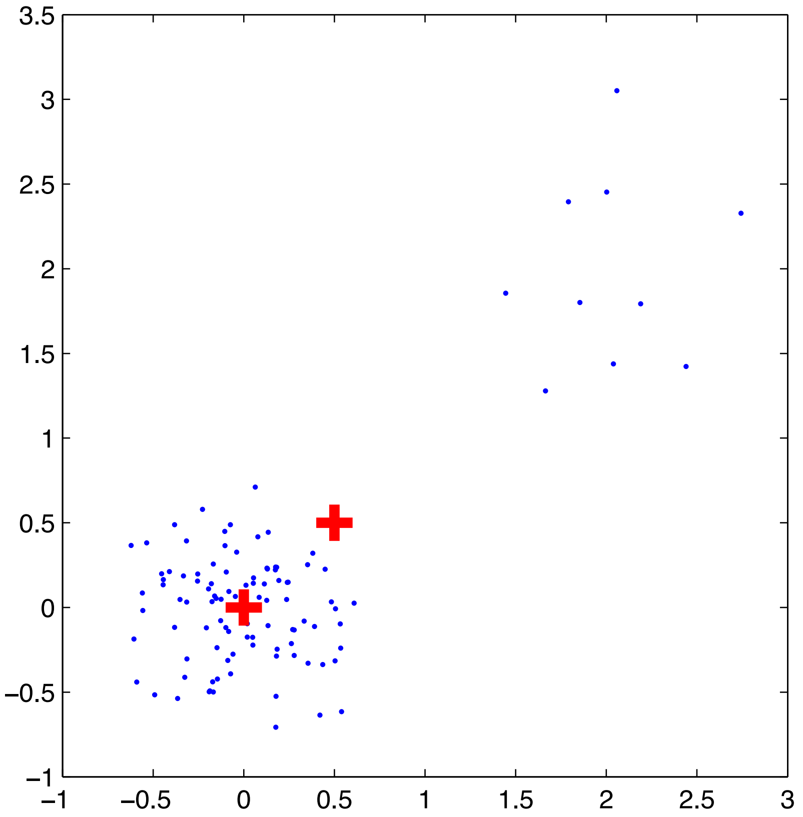


Figure 2: Initial data and cluster centers

Samples Questions

2.2 Lloyd's algorithm

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

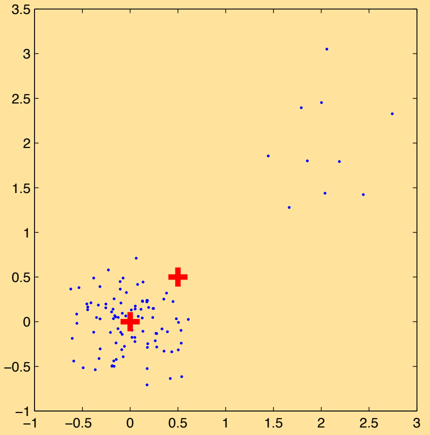
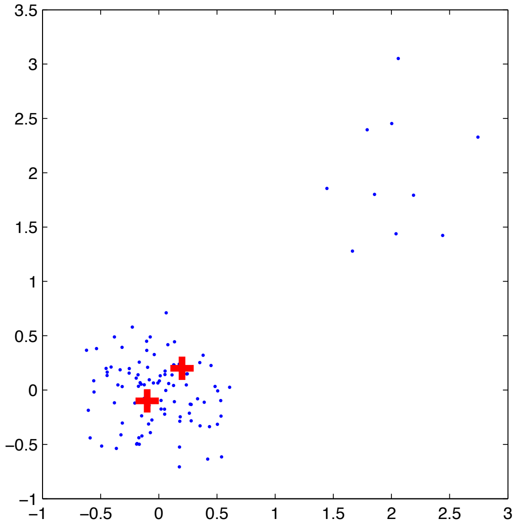
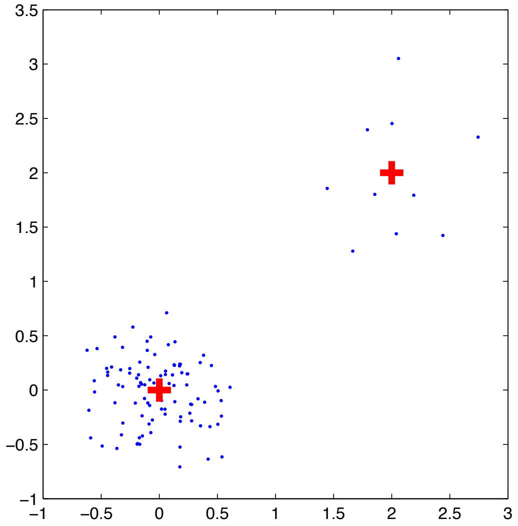
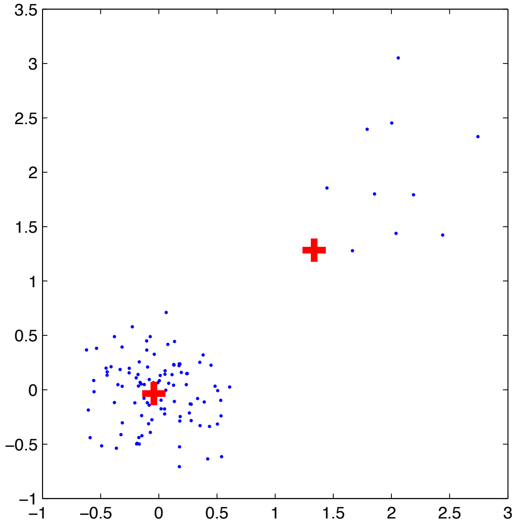
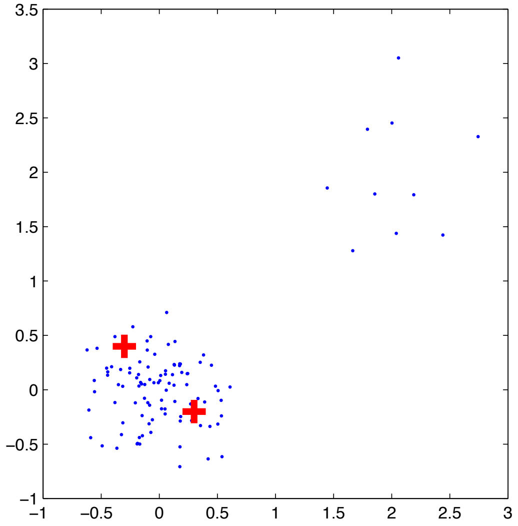


Figure 2: Initial data and cluster centers









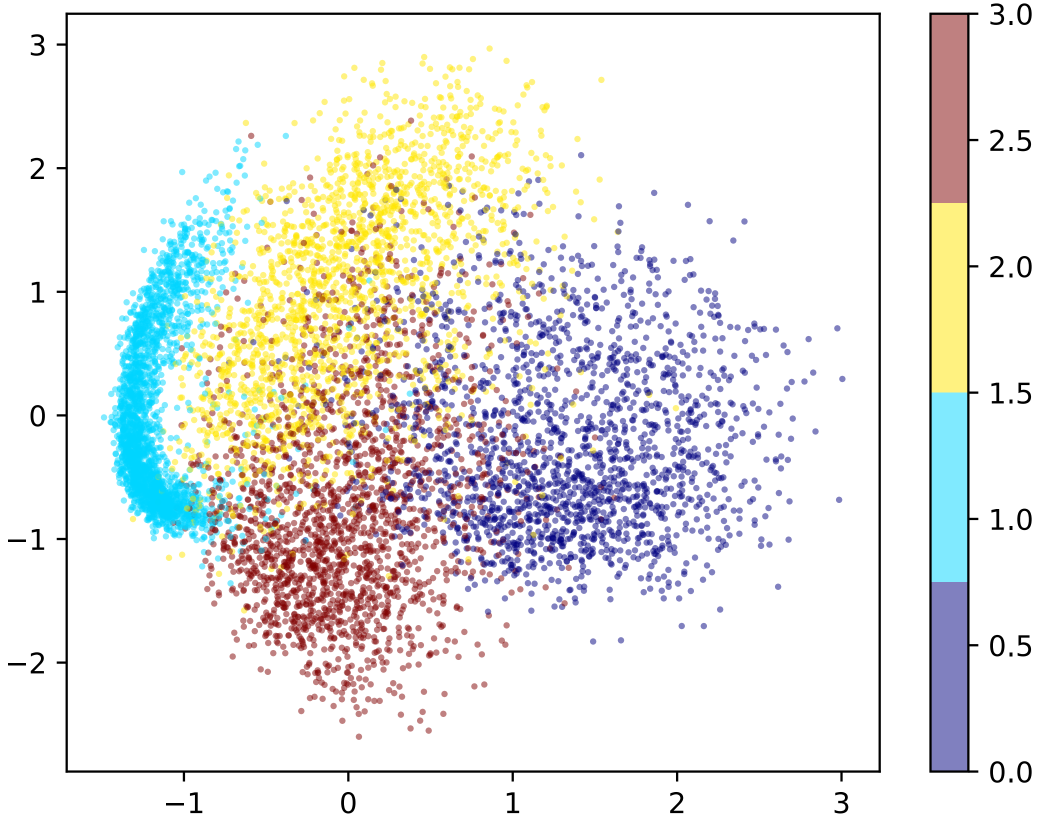
Shortcut Example

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

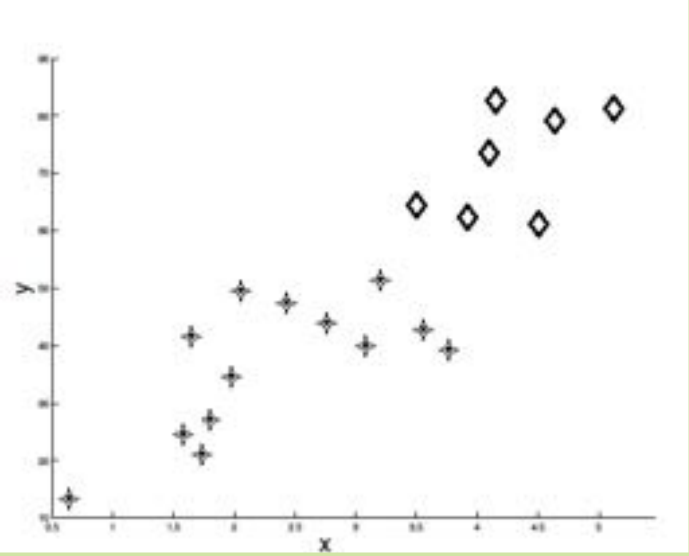
4 Principal Component Analysis [16 pts.]

(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:

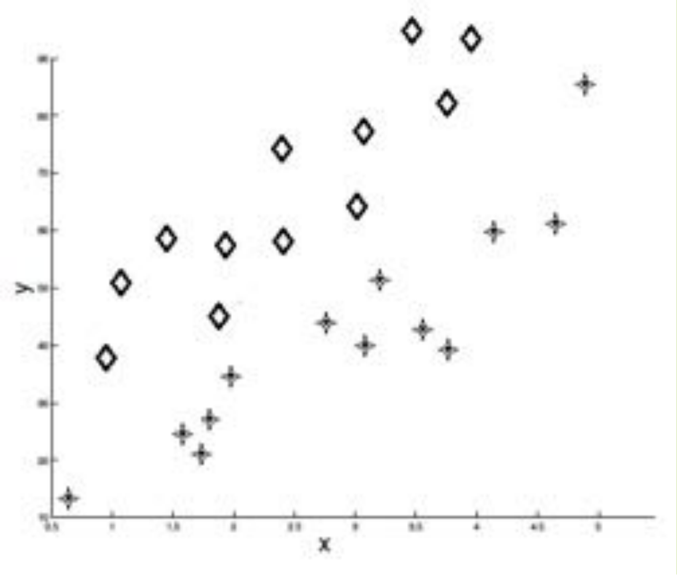
(i) [3 pt.] Draw all the principal components.

(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.

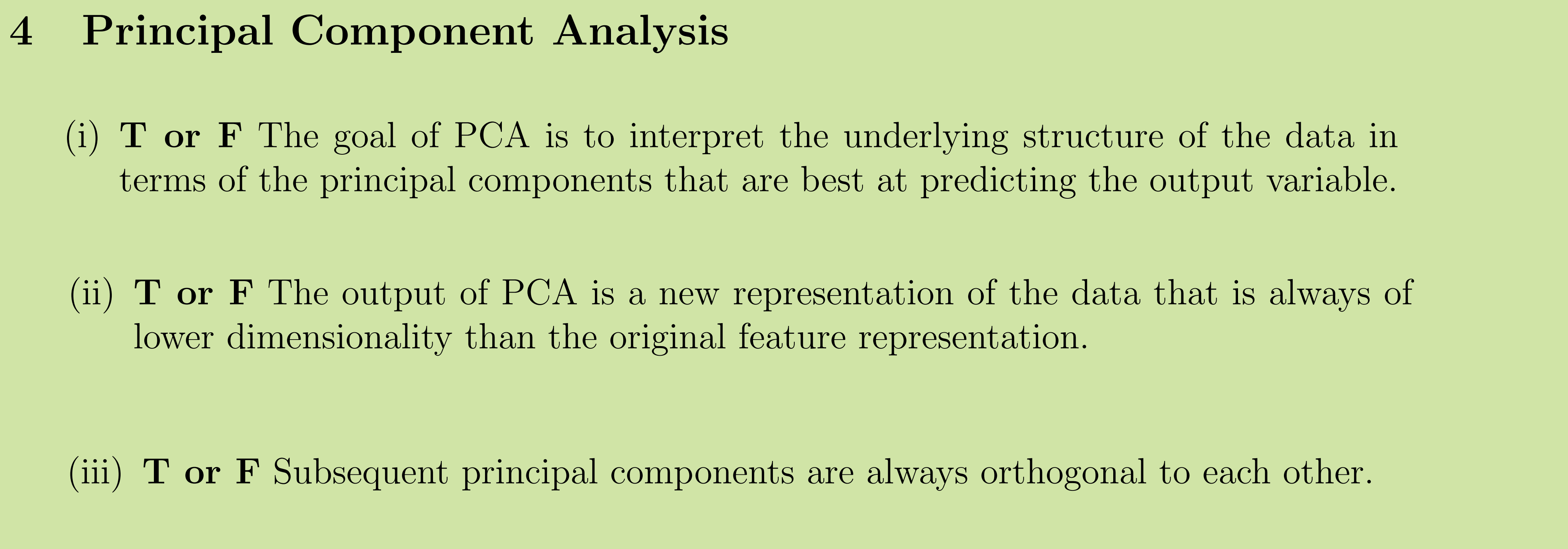
Dataset 1:



Dataset 2:



Sample Questions



4 Principal Component Analysis

(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.

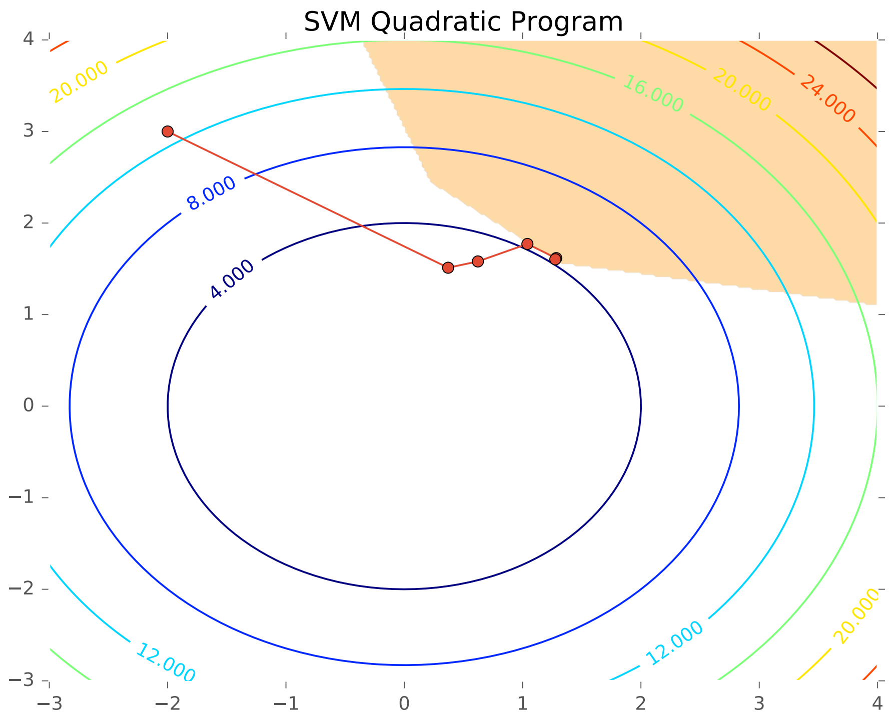
(ii) **T or F** The output of PCA is a new representation of the data that is always of lower dimensionality than the original feature representation.

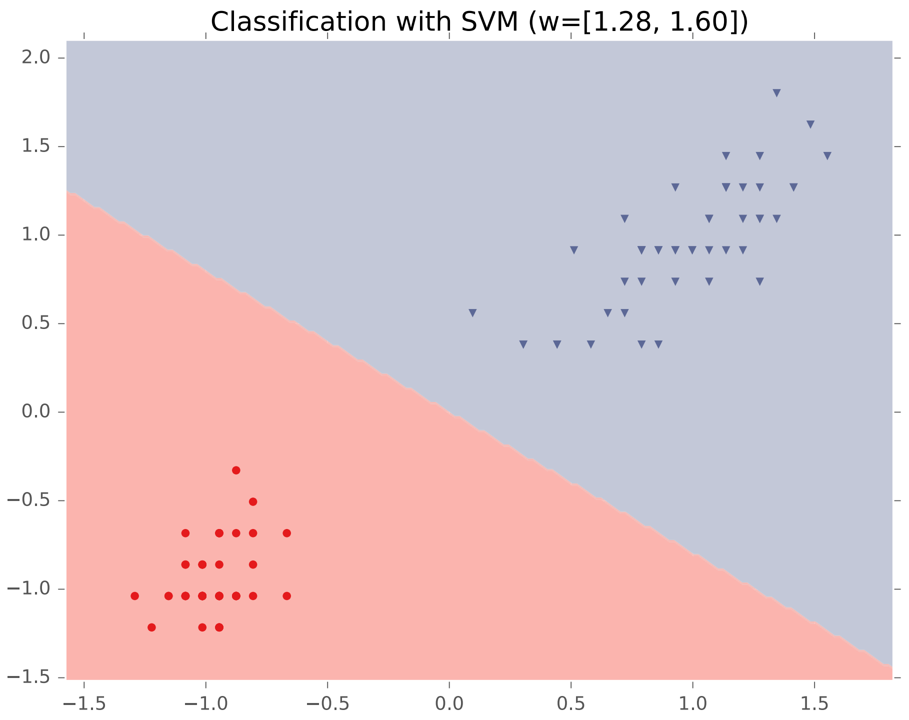
(iii) **T or F** Subsequent principal components are always orthogonal to each other.

Example: Building a Canal

https://www.flickr.com/photos/hereistom/10438848375

SVM QP


SVM Quadratic Program
4.000
8.000
12.000
16.000
20.000
24.000
−3
−2
−1
0
1
2
3
4


Classification with SVM (w=[1.28, 1.60])
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
0.0
0.5
1.0
1.5

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.

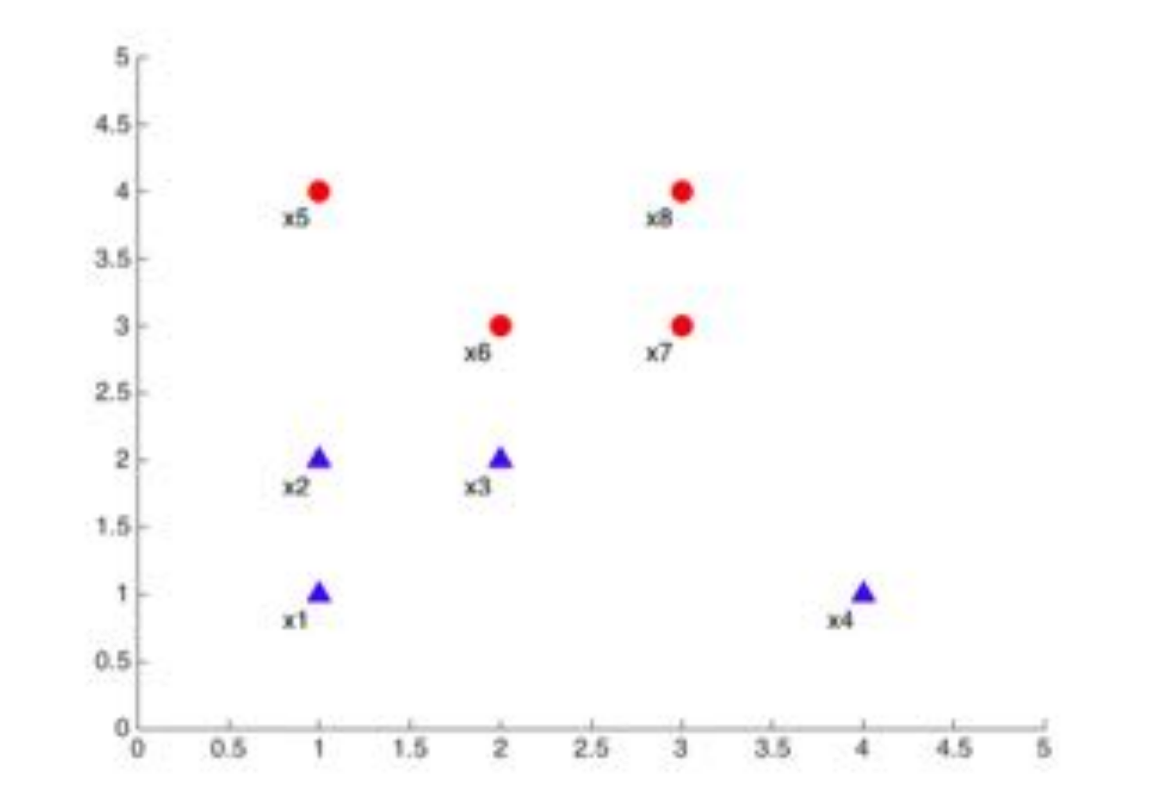


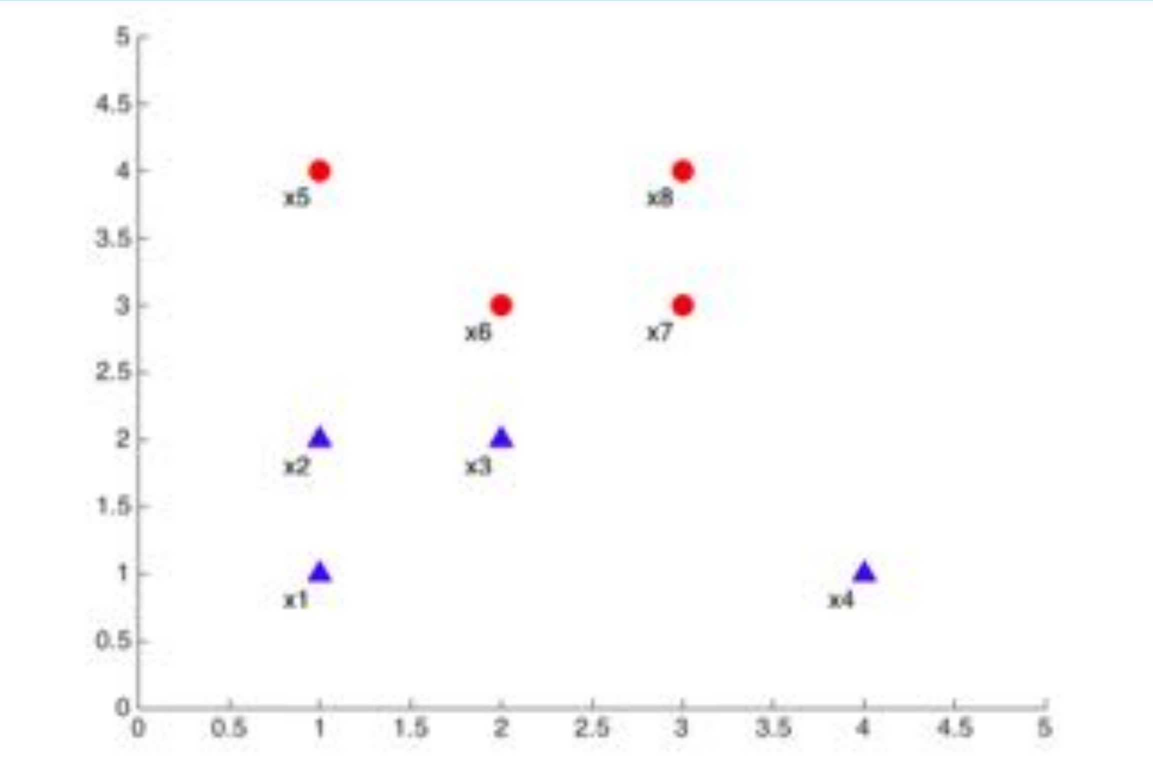
Figure 4: SVM toy dataset

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



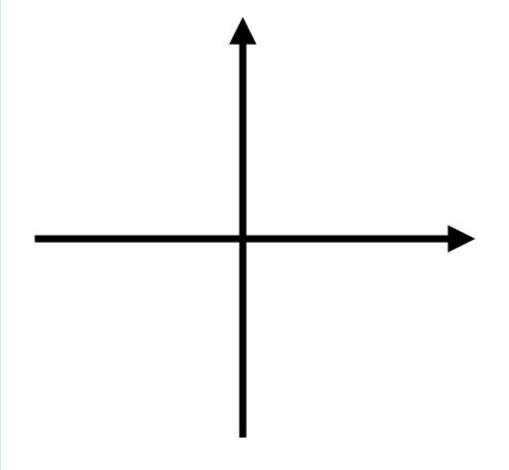
Sample Questions

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

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

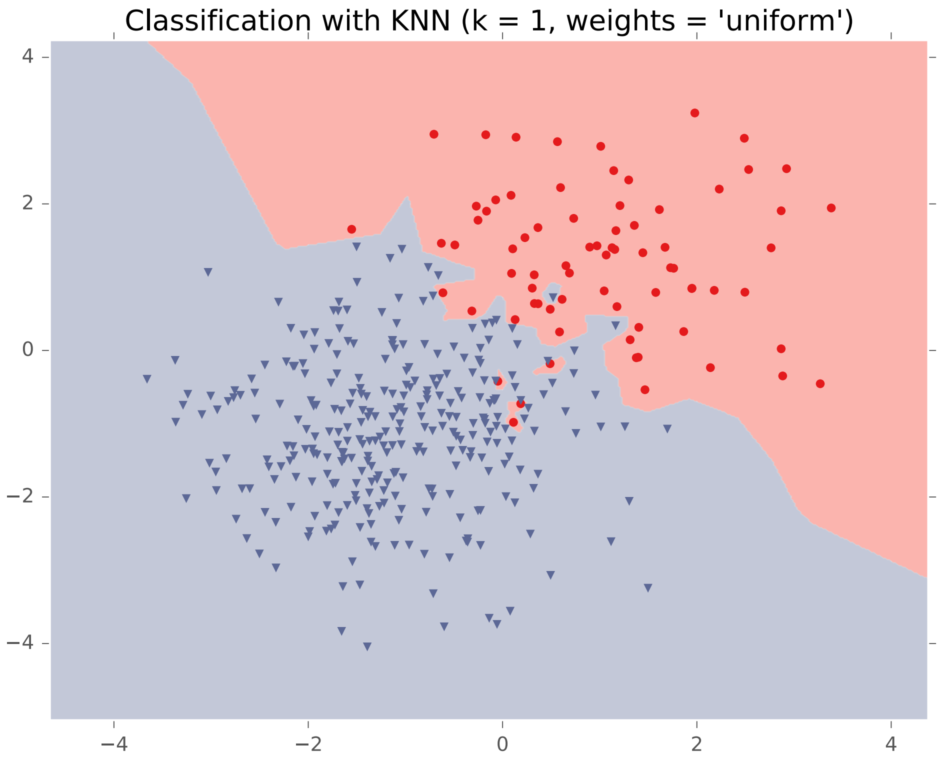
where (x_i, y_i) are training samples and $\mathbf{w}$ defines a linear decision boundary.

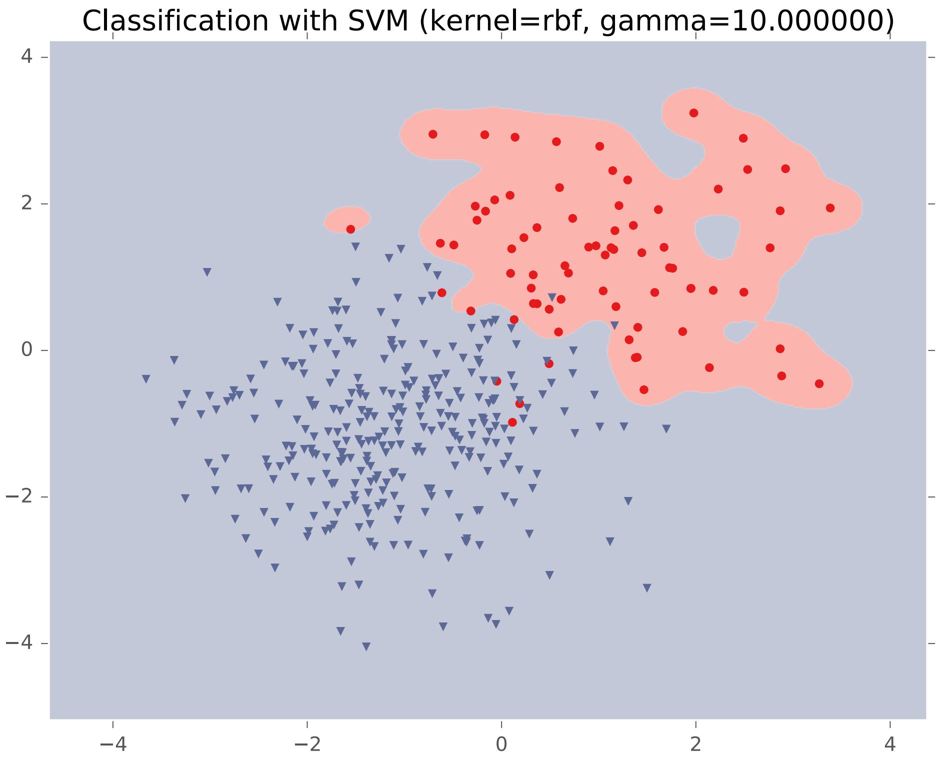
Derive a formula for ξ_i when the objective function achieves its minimum (No steps necessary). Note it is a function of $y_i \mathbf{w}^\top x_i$. Sketch a plot of ξ_i with $y_i \mathbf{w}^\top x_i$ on the x-axis and value of ξ_i on the y-axis. What is the name of this function?



RBF Kernel Example

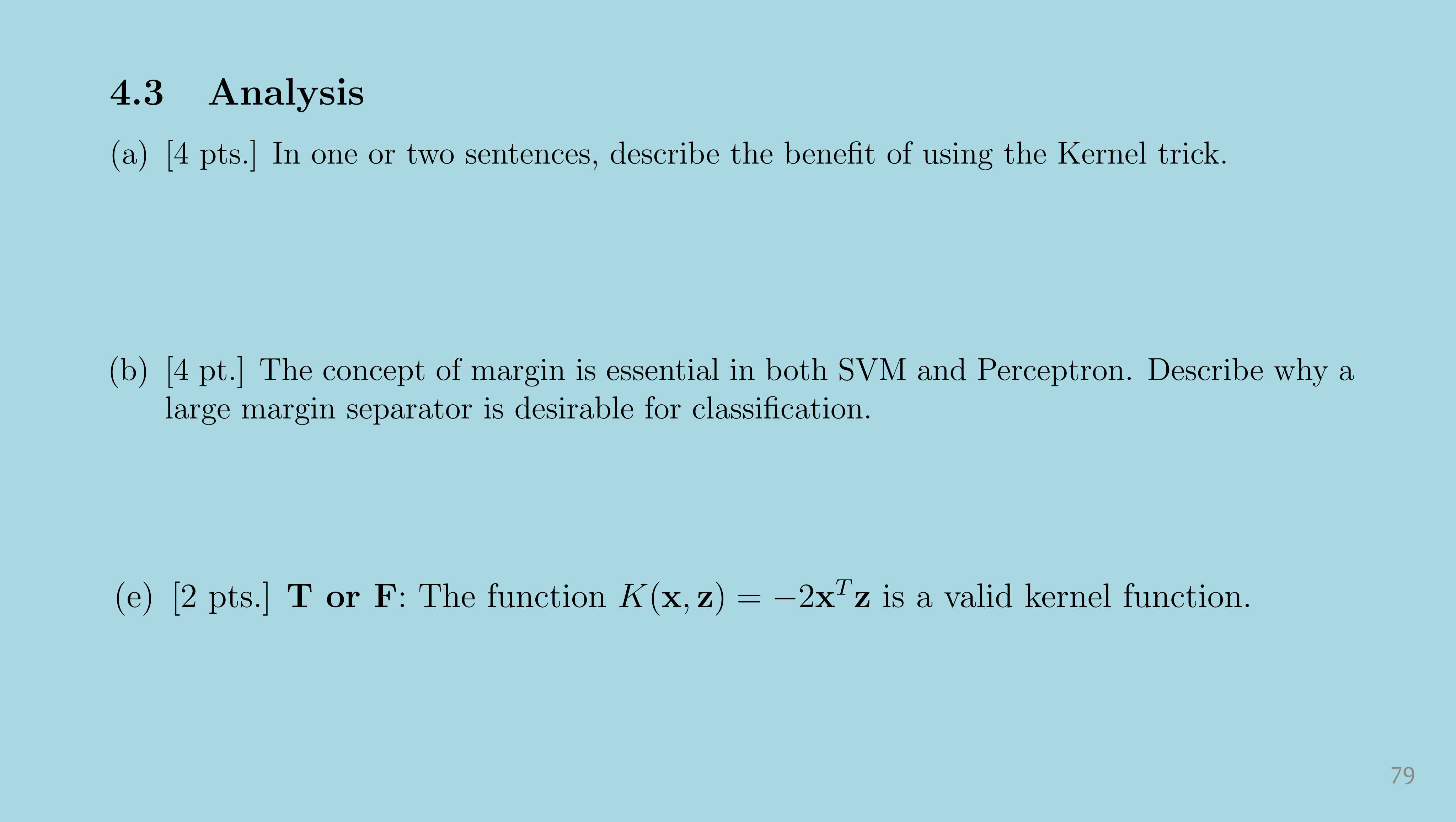
KNN vs. SVM


Classification with KNN (k = 1, weights = 'uniform')


Classification with SVM (kernel=rbf, gamma=10.000000)

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

Sample Questions



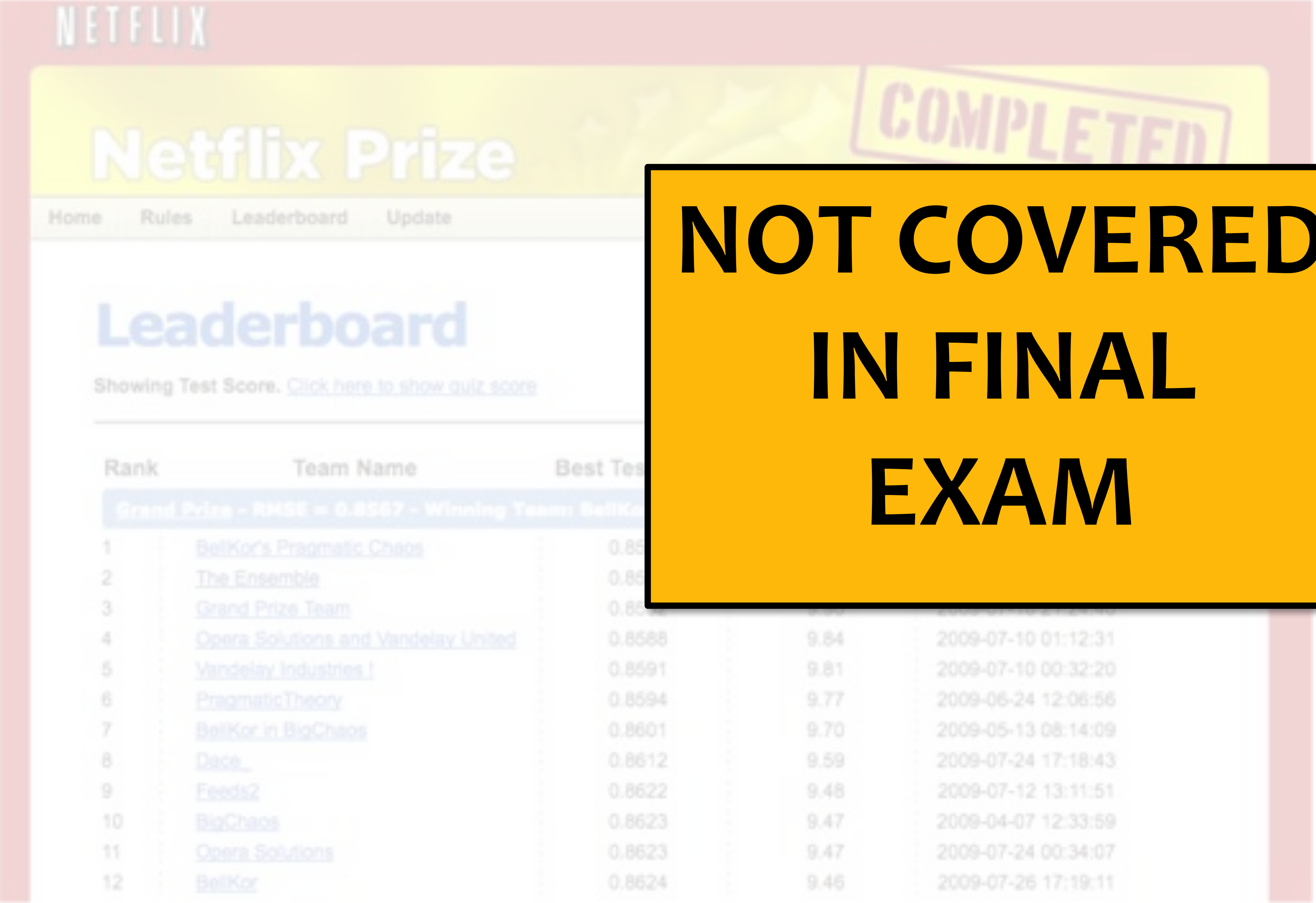
4.3 Analysis

(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



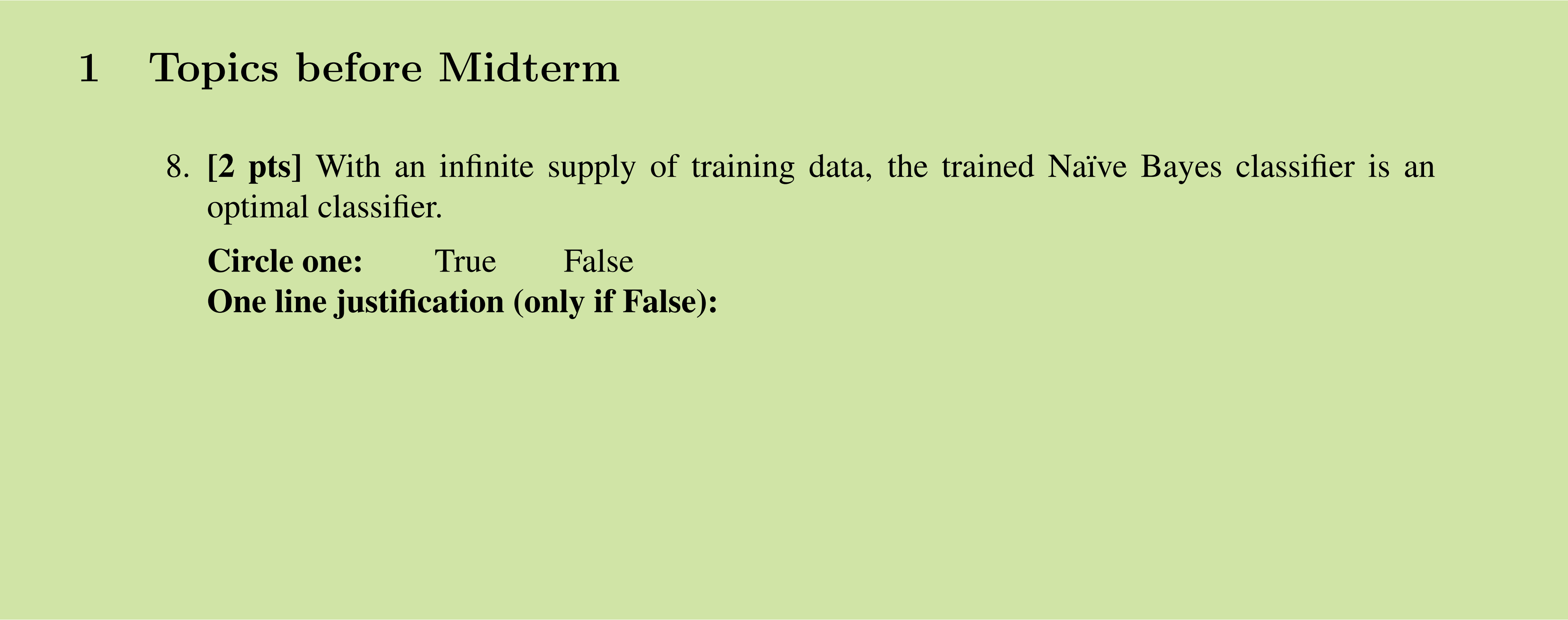
Leaderboard

Showing Test Score. Click here to show quiz score

Rank	Team Name	Best Tes		
Grand Prize - RMSE = 0.8567 - Winning Team: BellKo				
1	BellKor's Pragmatic Chaos	0.85		
2	The Ensemble	0.85		
3	Grand Prize Team	0.85		
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

NOT COVERED IN FINAL EXAM

Sample Questions



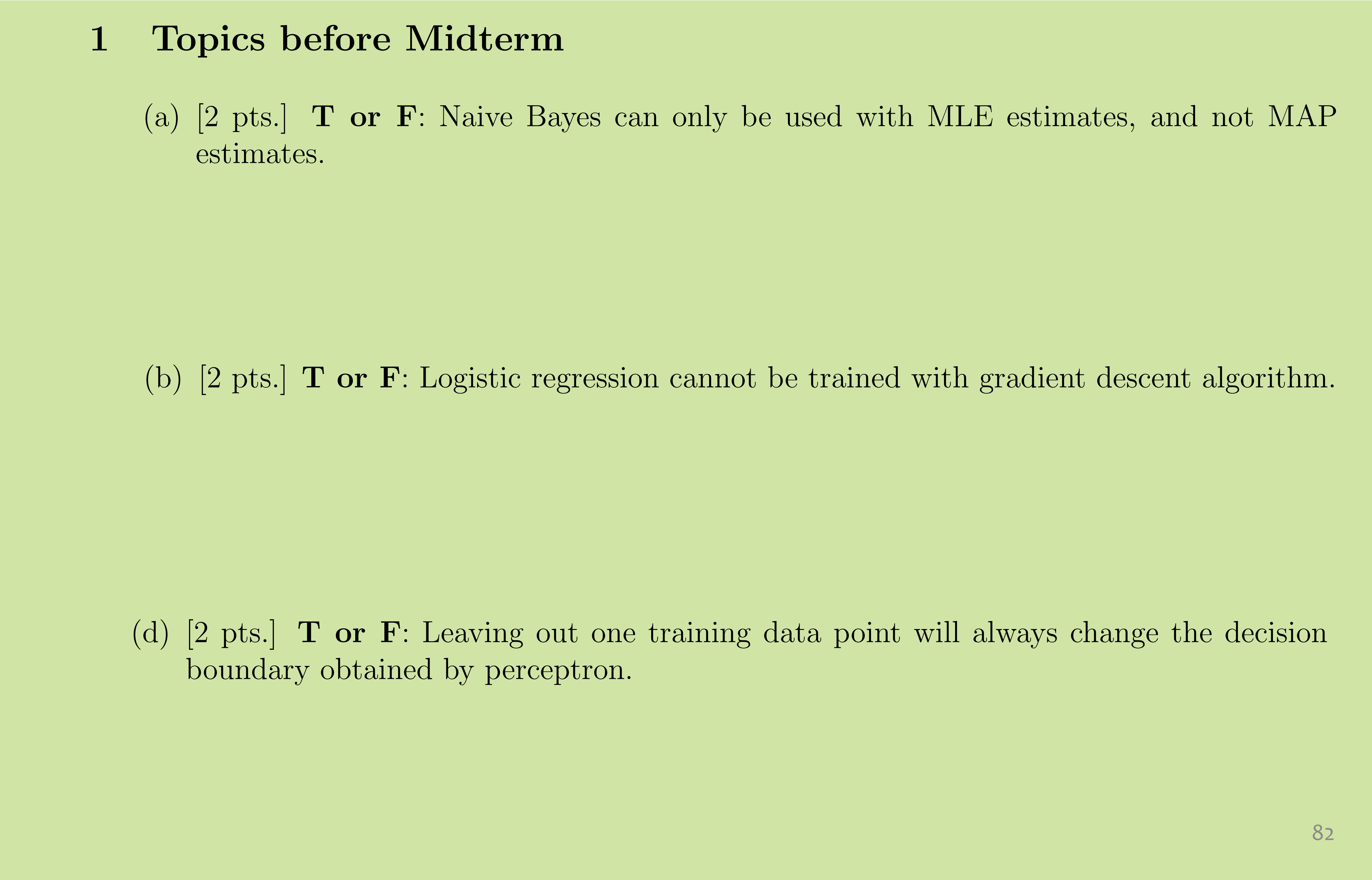
1 Topics before Midterm

8. **[2 pts]** With an infinite supply of training data, the trained Naïve Bayes classifier is an optimal classifier.

   Circle one: True False

   One line justification (only if False):

Sample Questions



1 Topics before Midterm

(a) [2 pts.] **T or F**: Naive Bayes can only be used with MLE estimates, and not MAP estimates.

(b) [2 pts.] **T or F**: Logistic regression cannot be trained with gradient descent algorithm.

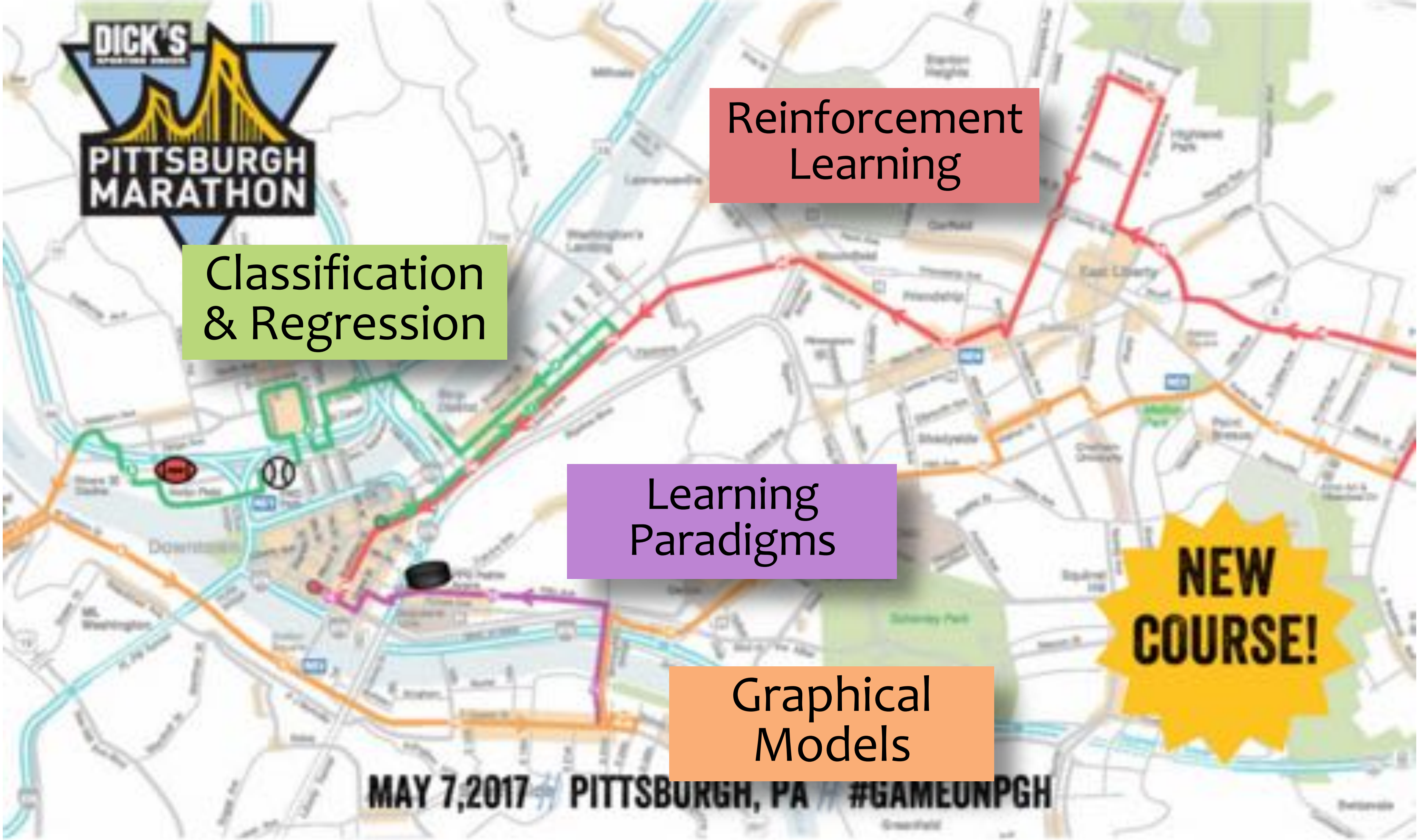
(d) [2 pts.] **T or F**: Leaving out one training data point will always change the decision boundary obtained by perceptron.

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:



Reinforcement Learning

Classification & Regression

Learning Paradigms

Graphical Models

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