10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Final Exam Review

Readings:
Murphy (all chapters)
Bishop (all chapters)
HTF (all chapters)
Mitchell (all chapters)

Matt Gormley
Lecture 29
May 3, 2016

Reminders

- Homework 9: Applications of ML
  - Release: Mon, Apr. 24
  - Due: Wed, May 3 at 11:59pm

- Final Exam (Evening Exam)
  - Mon, May 08 at 5:30pm – 8:30pm
  - See Piazza for details about location

Outline

1. Exam Logistics
2. Sample Questions
3. Overview

EXAM LOGISTICS

Final Exam

- **Time / Location**
  - **Time:** Evening Exam
    Mon, May 8 at 5:30pm – 8:30pm
  - **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**
  - 8-9 Sections
  - Format of questions:
    - Multiple choice
    - True / False (with justification)
    - Derivations
    - Short answers
    - Interpreting figures
  - 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 will post them shortly
    - Disclaimer: This year's 10-601 is not the same as prior offerings
  - Review this year's homework problems
  - Attend the **Mock Final Exam**
    - Thu, May 4, 6:30pm
    - Section A should go to PH100
    - Section B and C should go to DH2210
    - Disclaimer: The Mock will be much shorter and not exhaustive, but great practice!

Final Exam

- **How to Prepare**
  - Attend the final recitation session:
    Tue, Dec. 6th at 5:30pm
  - Review prior year's exams and solutions
    (we will post them)
  - Review this year's homework problems
  - Flip through the "What you should know" points
    (see 'More' links on 'Schedule' page of course
    website)

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**
  - 10-20% of material comes from topics covered **before** the midterm exam
  - 80-90% of material comes from topics covered **after** the midterm exam

Topics covered **before** Midterm

- Foundations
  - Probability
  - MLE, MAP
  - Optimization
- Classifiers
  - KNN
  - Naïve Bayes
  - Logistic Regression
  - Perceptron
  - SVM
- Regression
  - Linear Regression
- Important Concepts
  - Kernels
  - Regularization and Overfitting
  - Experimental Design

Topics covered **after** Midterm

- Unsupervised Learning
  - K-means / Lloyd's method
  - PCA
  - EM / GMMs
- Neural Networks
  - Feedforward Neural Nets
  - Basic architectures
  - Backpropagation
  - CNNs
- Graphical Models
  - Bayesian Networks
  - HMMs
  - Learning and Inference
- Learning Theory
  - Statistical Estimation (covered right before midterm)
  - PAC Learning
- Other Learning Paradigms
  - Matrix Factorization
  - Reinforcement Learning
  - Information Theory

SAMPLE QUESTIONS

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.

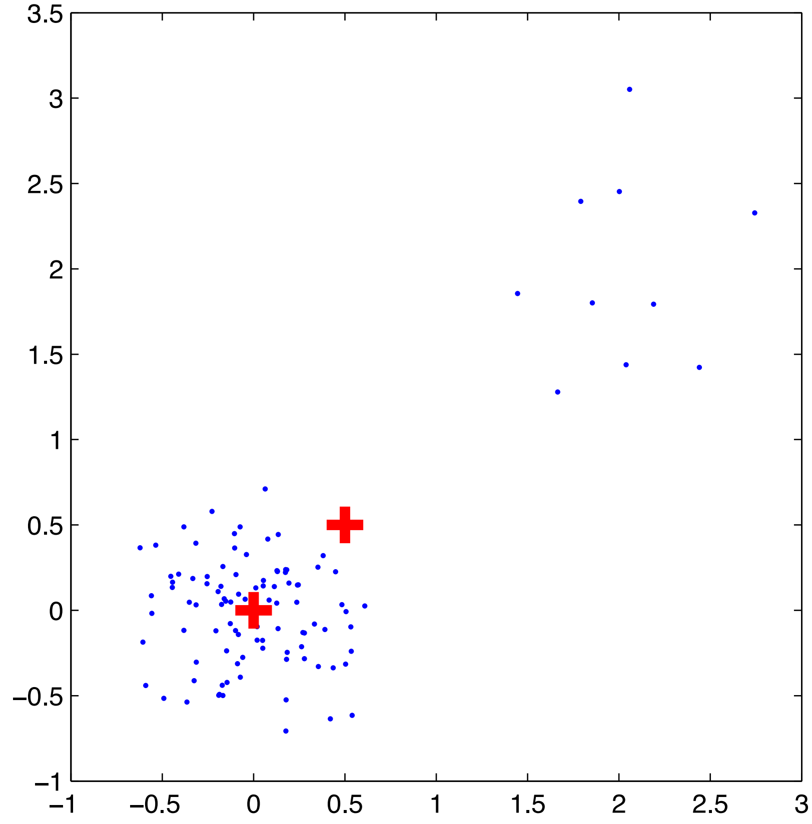


Figure 2: Initial data and cluster centers

Samples Questions

2.2 Lloyd's algorithm

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

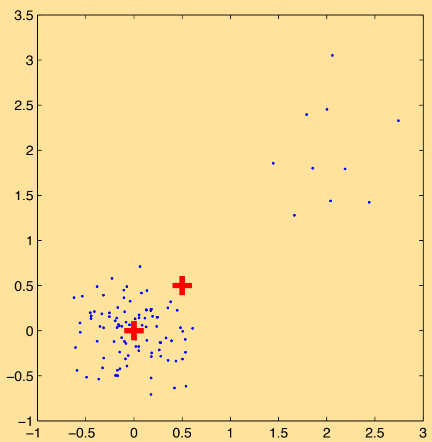
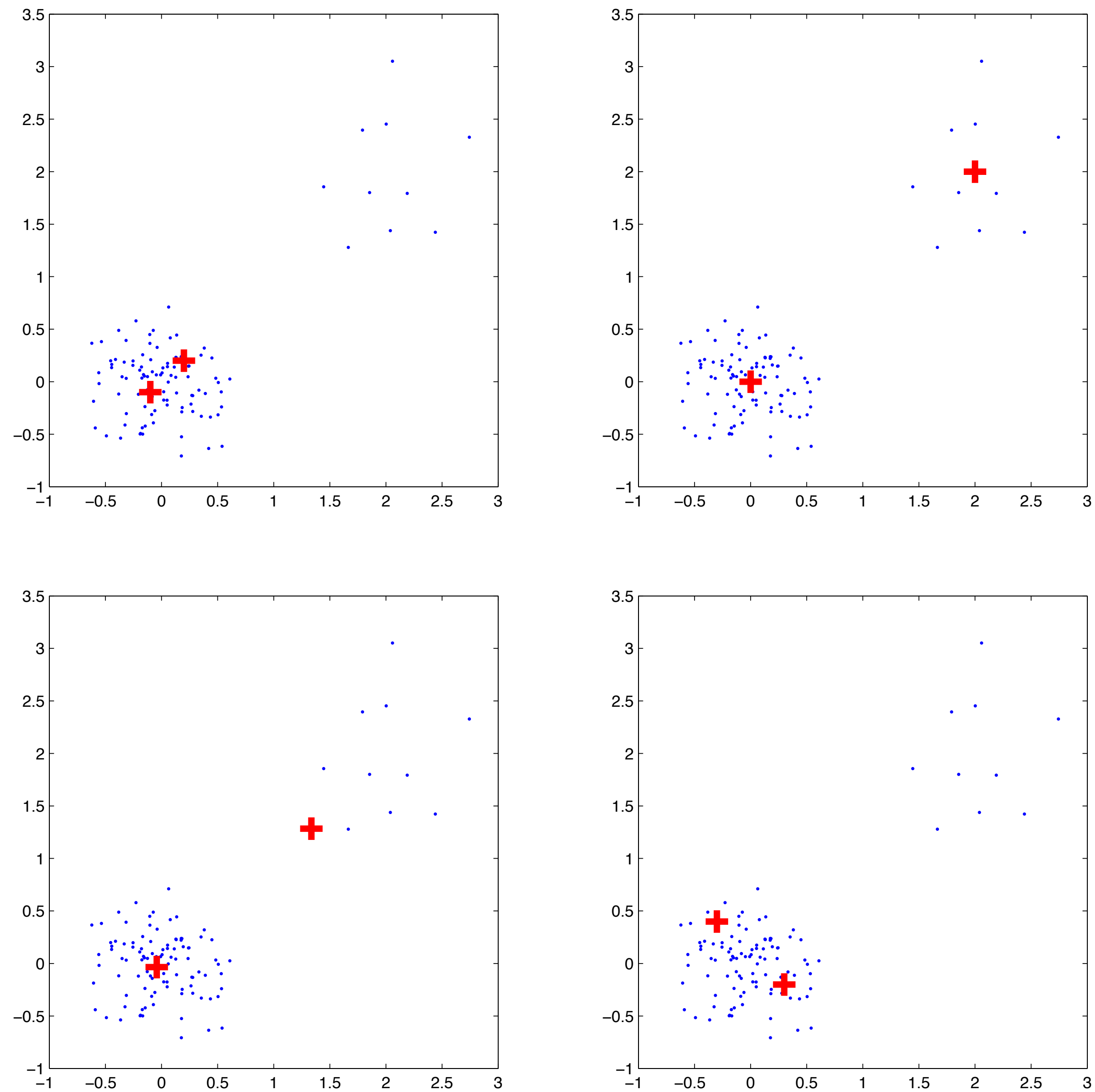


Figure 2: Initial data and cluster centers



Sample Questions

Question 4: Expectation Maximization

Given a set of observed variables X, a set of latent variables Z, and a set of model parameters with the current estimate being θ, a single iteration of the EM algorithm updates the parameters estimate θ as follows:

$$\theta \leftarrow \arg\max_{\theta'} Q(\theta'|\theta) \equiv \mathbb{E}_{P(Z|X,\theta)}[\log P(X, Z|\theta')]$$

where $\log P(X, Z|\theta') = \log \prod_{i=1}^{n} P(X_i, Z_i|\theta')$ is known as the *complete log likelihood* of the data.

(a) [2 pts] True or False: In the case of fully observed data, i.e. when Z is an empty set, the EM algorithm reduces to a maximum likelihood estimate.

(b) [2 pts] True or False: Since the EM algorithm guarantees that the value of its objective function will increase on each iteration, it is guaranteed to eventually reach a global maximum.

Sample Questions

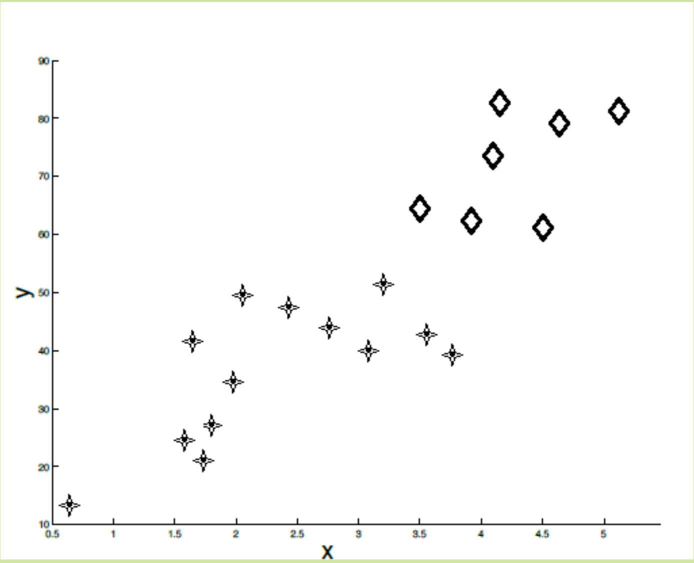
4 Principal Component Analysis [16 pts.]

(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:

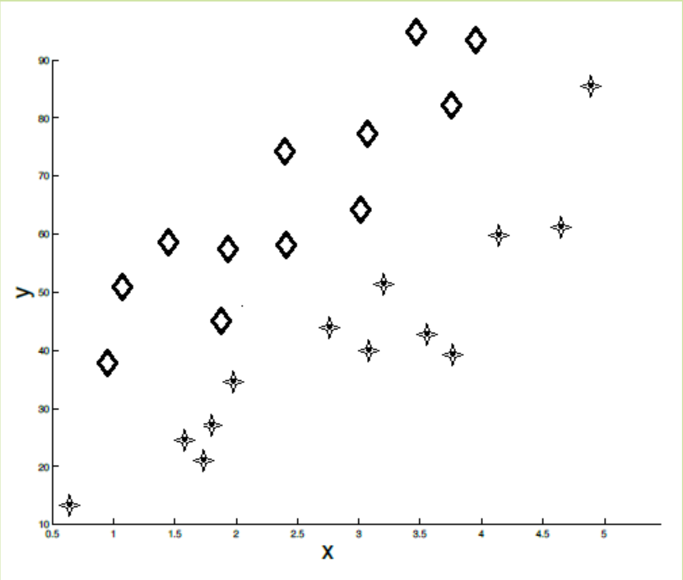
   (i) [3 pt.] Draw all the principal components.

   (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.

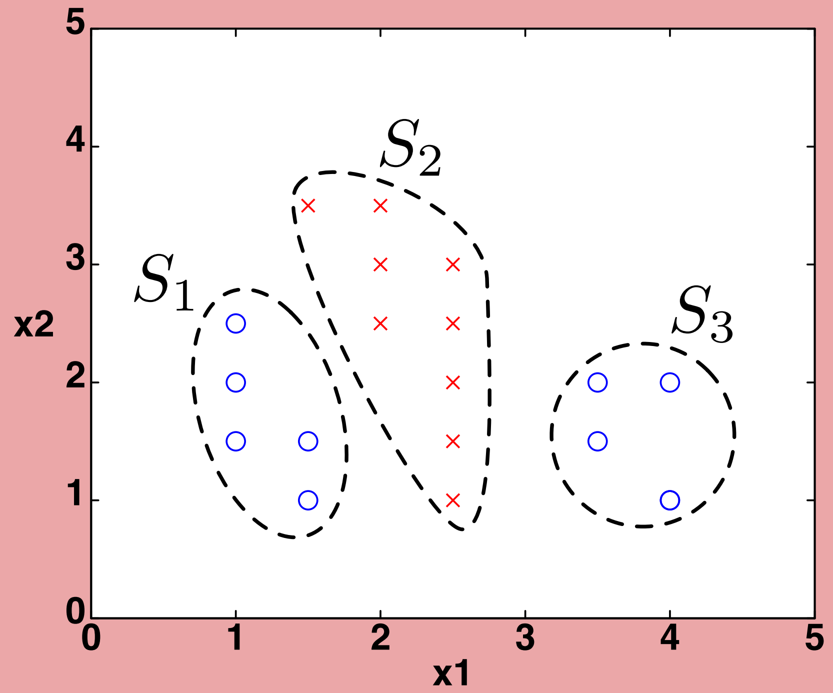
(ii) **T or F** The output of PCA is a new representation of the data that is always of lower dimensionality than the original feature representation.

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

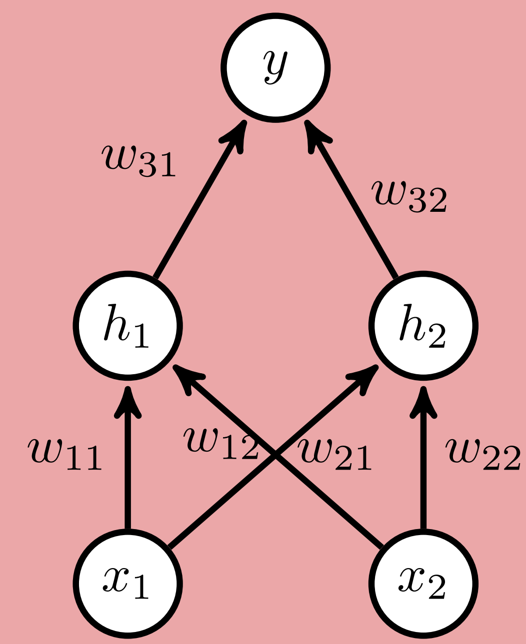
Sample Questions

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.

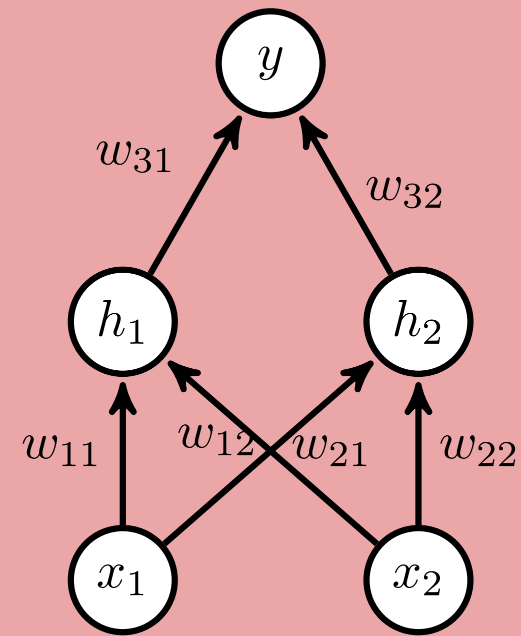


(b) The neural network architecture

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

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

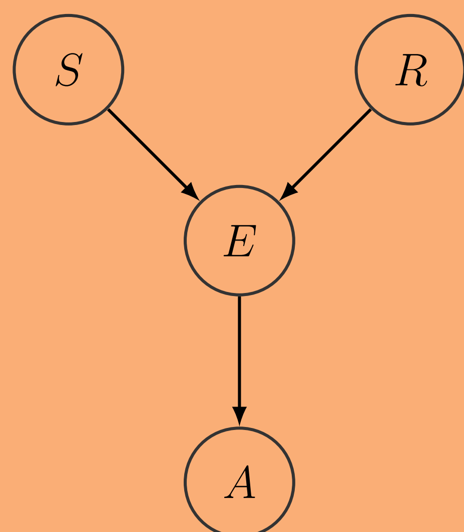


Figure 5: Directed graphical model for problem 5.

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

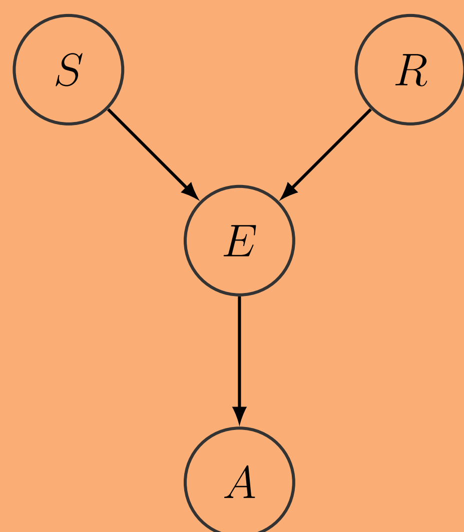


Figure 5: Directed graphical model for problem 5.

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

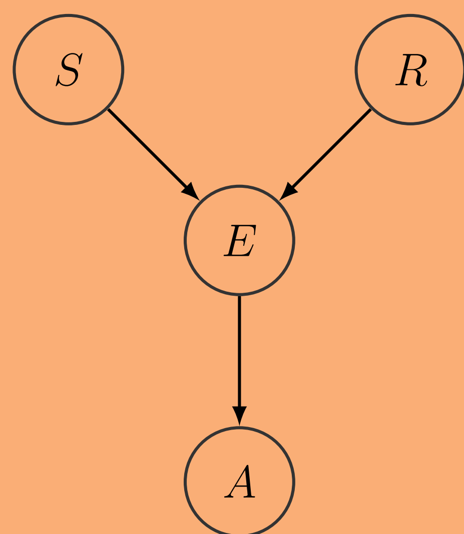


Figure 5: Directed graphical model for problem 5.

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.

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.

Sample Questions

1 Topics before Midterm

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

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):

OVERVIEW

Whiteboard

- Overview #1: Learning Paradigms
- Overview #2: Recipe for ML