10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Final Exam Review

Matt Gormley
Lecture 31
May 2, 2018

Reminders

- **Homework 9: Learning Paradigms**
  - **Out: Sat, Apr 28**
  - **Due: Fri, May 4 at 11:59pm**

Outline

1. Exam Logistics
2. Sample Questions
3. Overview

EXAM LOGISTICS

Final Exam

- **Time / Location**
  - **Time:** Evening Exam
    Mon, May 14 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
  - Review this year's homework problems
  - 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**
  - ~20% of material comes from topics covered **before** the midterm exam
  - ~80% of material comes from topics covered **after** the midterm exam

Topics covered **before** Midterm

- Foundations
  - Probability, Linear Algebra, Geometry, Calculus
  - MLE
  - Optimization
- Important Concepts
  - Regularization and Overfitting
  - Experimental Design
- Classifiers
  - Decision Tree
  - KNN
  - Perceptron
  - Logistic Regression
- Regression
  - Linear Regression
- Feature Learning
  - Neural Networks
  - Basic NN Architectures
  - Backpropagation
- Learning Theory
  - PAC Learning

Topics covered **after** Midterm

- Learning Theory
  - PAC Learning
- Generative Models
  - Generative vs. Discriminative
  - MLE / MAP
  - Naïve Bayes
- 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 / AdaBoost

Material Covered **After** Midterm Exam

SAMPLE QUESTIONS

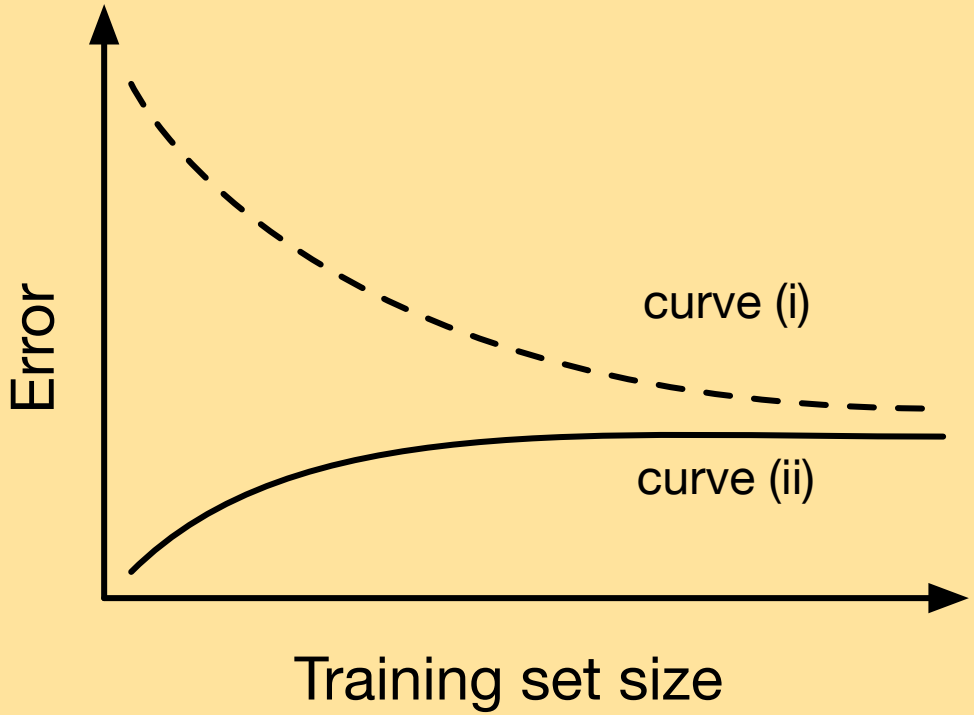
Samples Questions

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:** $P(\texttt{height}|\texttt{sex},\texttt{hair}) = P(\texttt{height}|\texttt{sex})$.

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

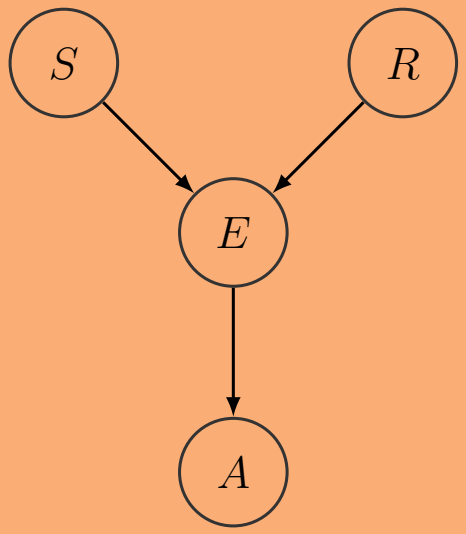


Figure 5: Directed graphical model for problem 5.

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

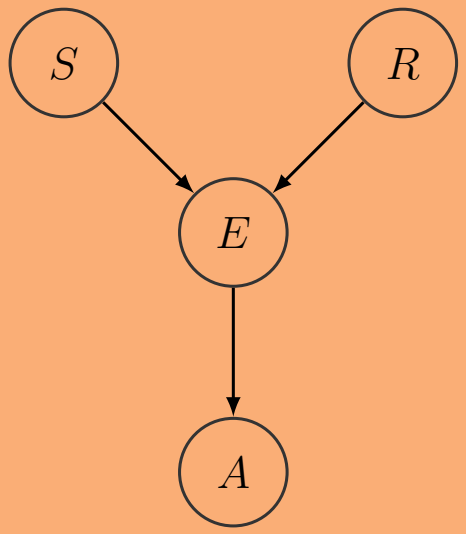


Figure 5: Directed graphical model for problem 5.

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

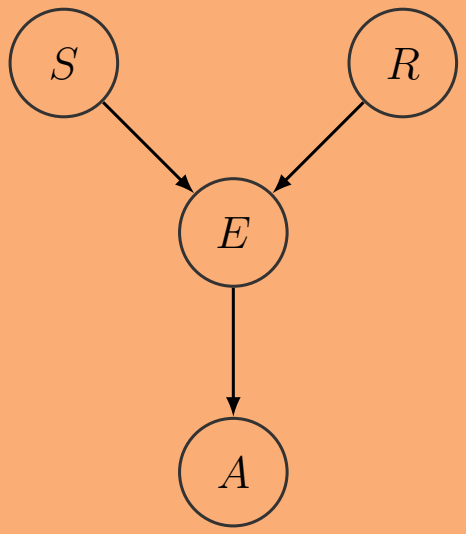


Figure 5: Directed graphical model for problem 5.

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

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.

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.

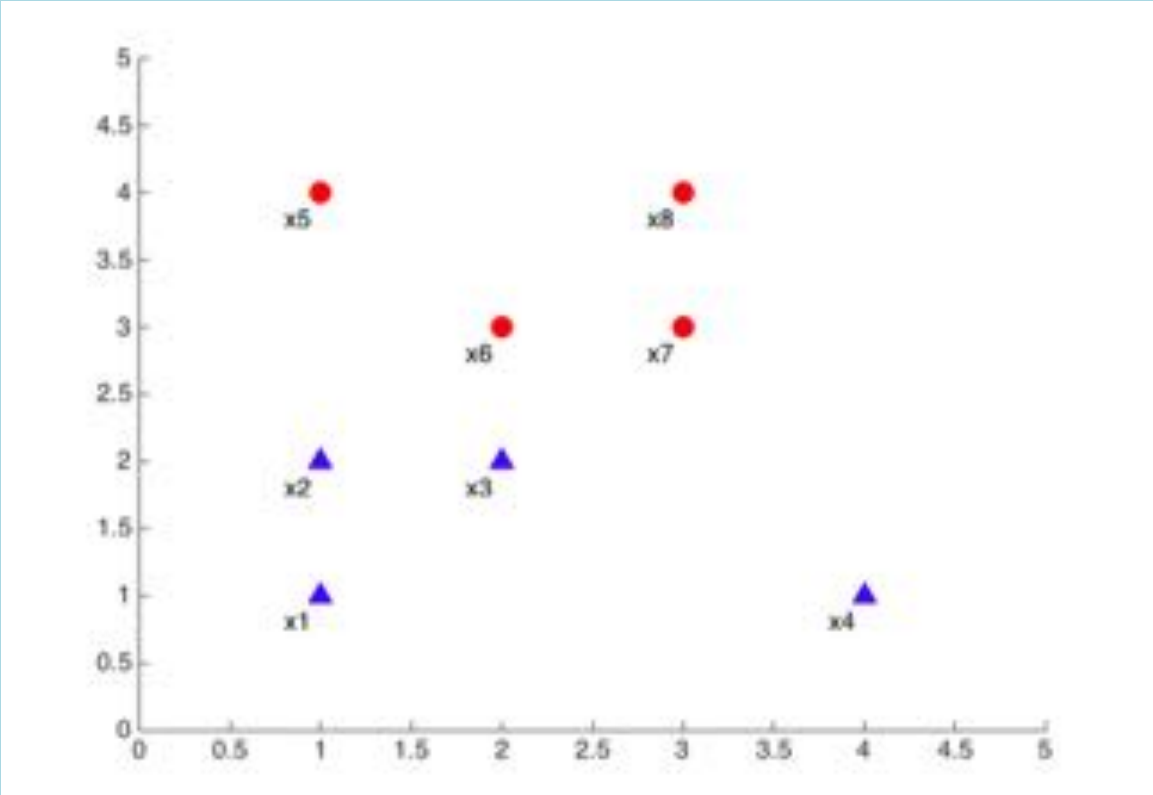


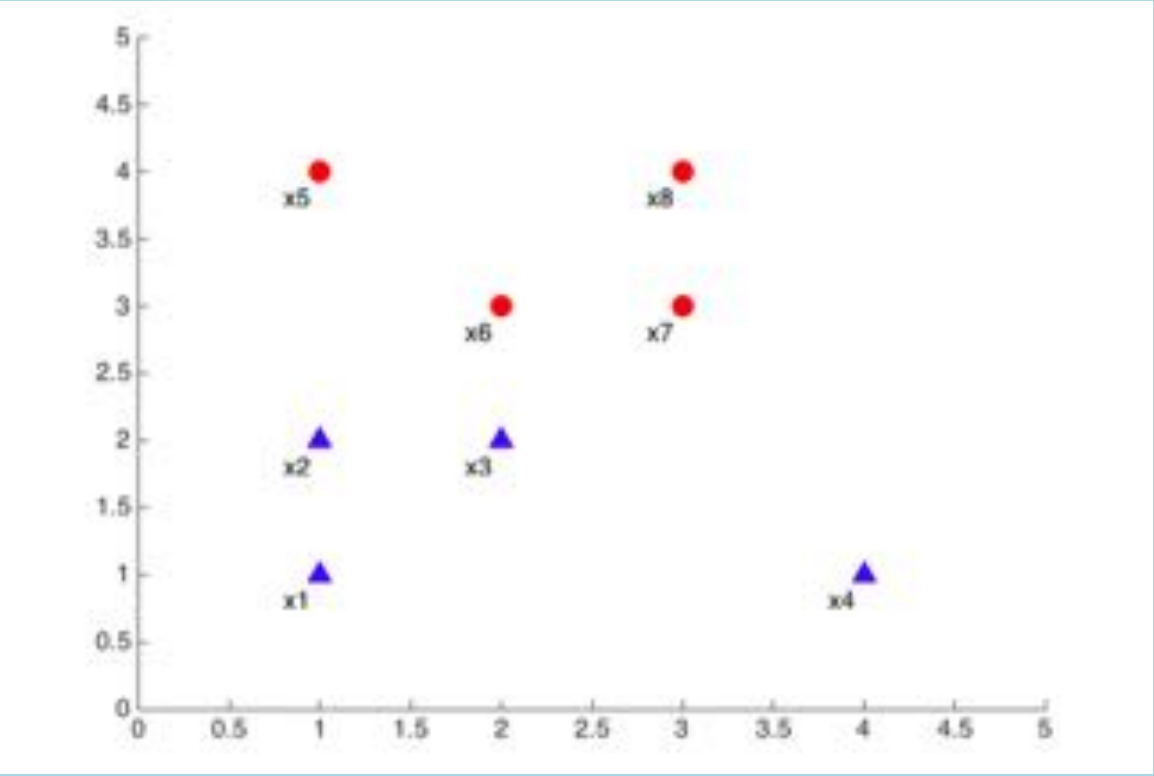
Figure 4: SVM toy dataset

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



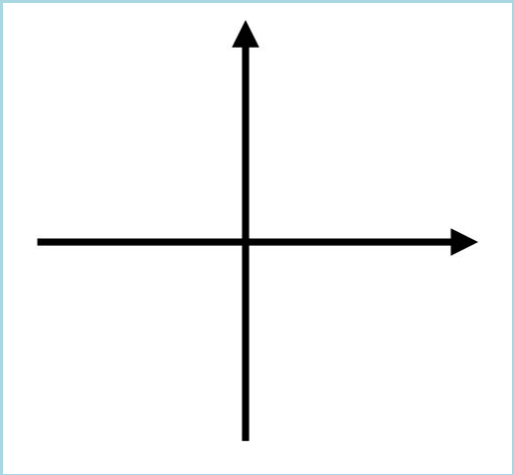
Sample Questions

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?



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.

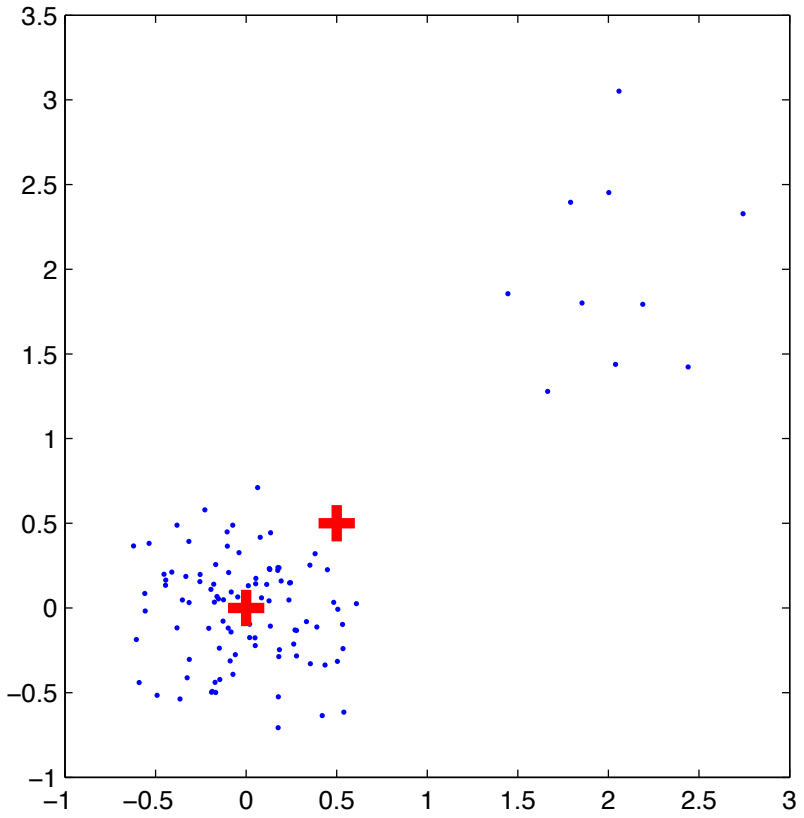


Figure 2: Initial data and cluster centers

Samples Questions

2.2 Lloyd's algorithm

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

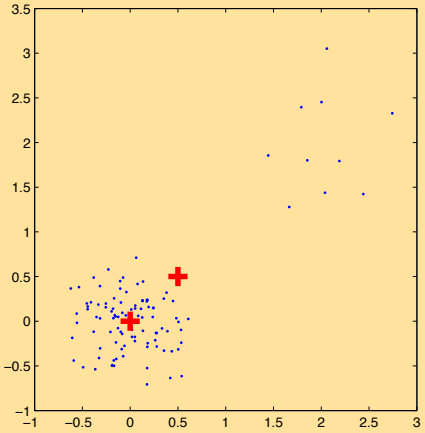
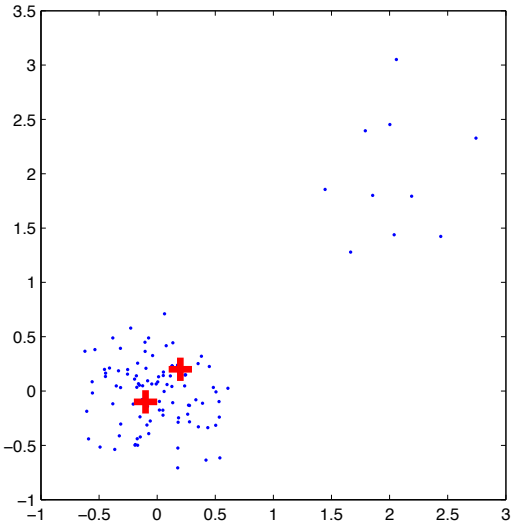
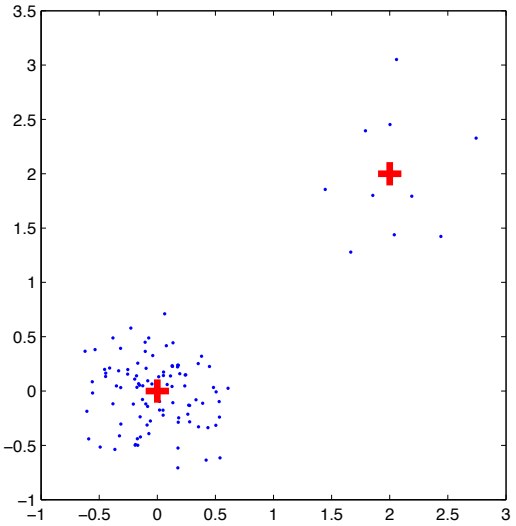
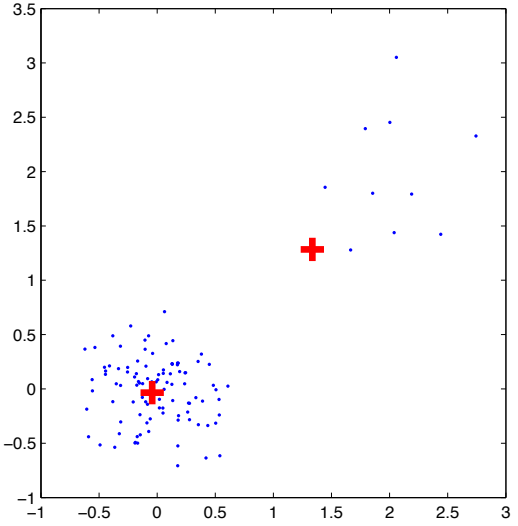
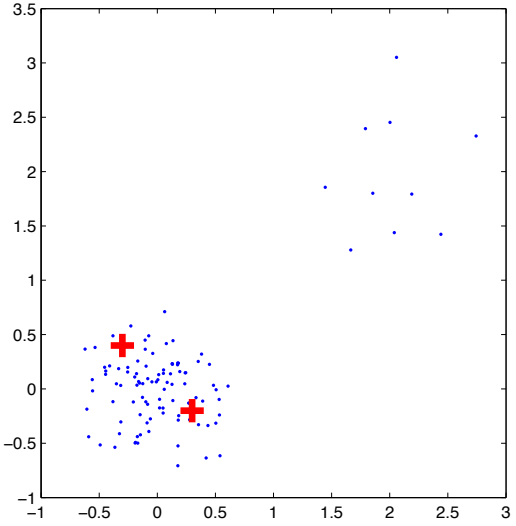


Figure 2: Initial data and cluster centers









Sample Questions

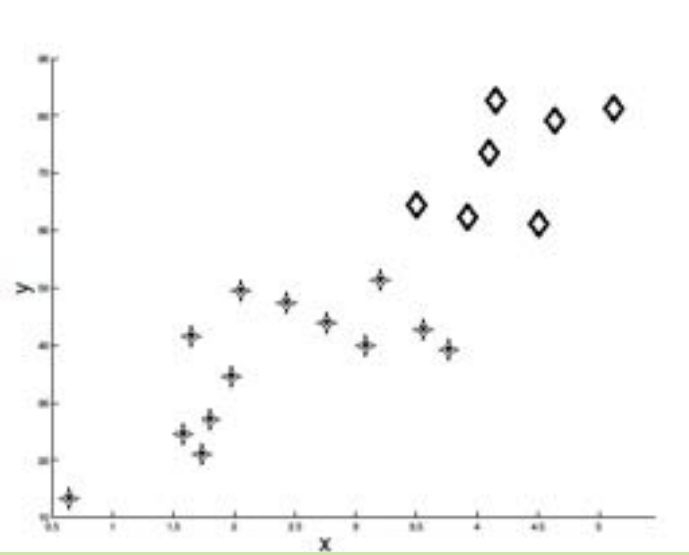
4 Principal Component Analysis [16 pts.]

(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:

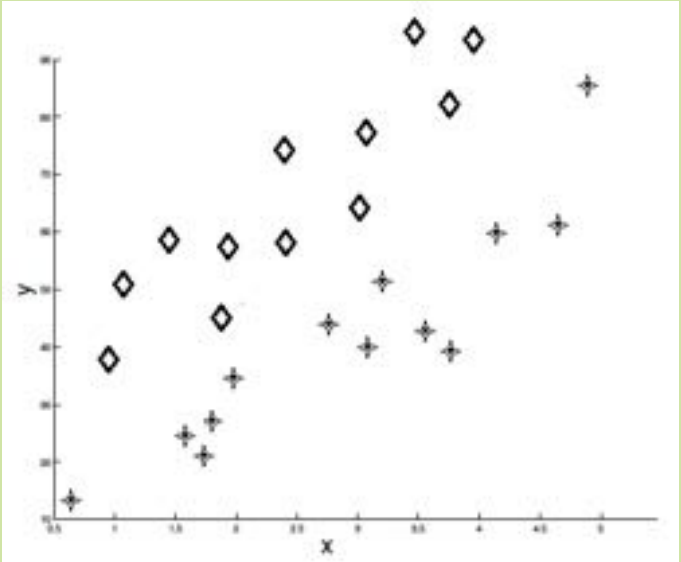
(i) [3 pt.] Draw all the principal components.

(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

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.