New York University Tandon School of Engineering
Computer Science and Engineering

Midterm Exam Sample Problems

Always, Sometimes, Never.

Indicate whether each of the following statements is ALWAYS true, SOMETIMES true, or NEVER true. For full credit, provide a short justification or example to explain your choice.

(a) For random events $p(x \mid y) < p(x, y)$.

ALWAYS SOMETIMES NEVER

(b) Consider a loss function L. If $\nabla L(\boldsymbol{\beta}) = 0$, then $\boldsymbol{\beta}$ is a minimizer of L.

ALWAYS SOMETIMES NEVER

(c) The empirical risk of a model is lower than the population risk.

ALWAYS SOMETIMES NEVER

(d) You use gradient descent to find parameters $\boldsymbol{\beta}_{GD}$ for a multiple linear regression problem under ℓ_2 loss: $L(\boldsymbol{\beta}) = \|X\boldsymbol{\beta} - \mathbf{y}\|_2^2$. You are short on time, so you only run gradient descent for 10 iterations. Your friend finds parameters $\boldsymbol{\beta}_M$ using the equation $\boldsymbol{\beta}_M = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. Is $L(\boldsymbol{\beta}_m) \leq L(\boldsymbol{\beta}_{GD})$?

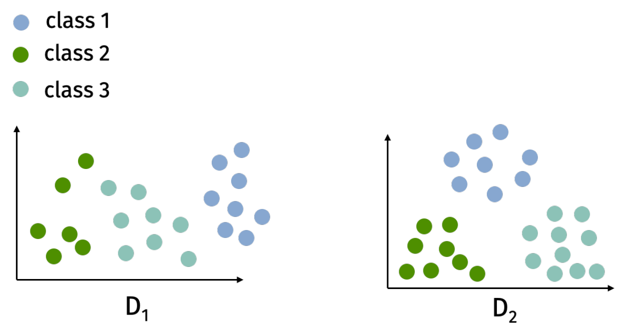
ALWAYS SOMETIMES NEVER

(e) Does $\boldsymbol{\beta}_M$ achieve better population risk than $\boldsymbol{\beta}_{GD}$?

ALWAYS SOMETIMES NEVER

(f) Consider a binary classification problem with negative class 0 and positive class 1. Consider two different linear classifiers that make predictions using the equation: (a) $\mathbb{1}[\langle \mathbf{x}, \boldsymbol{\beta} \rangle > 0]$ or (b) $\mathbb{1}[\langle \mathbf{x}, \boldsymbol{\beta} \rangle > 1]$, where $\boldsymbol{\beta}$ is a fixed parameter vector. Does classifier (a) have higher precision than classifier (b)?

ALWAYS SOMETIMES NEVER



Short Answer

Respond to each of the following questions using just a few words.

You are trying to develop a machine learning algorithm for classifying data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ into categories $1, \ldots, q$. You have decided to use linear classification for the problem.

(a) You know you can find a good linear classifier for binary classification (dividing into $q = 2$ classes) using logistic regression. You are considering using either the one-vs-all or one-vs-one approach to adapting this approach to the multiclass problem. In a few sort sentences describe why you might use one over the other.

(b) Your coworker suggests the following alternative approach: let's try to learn a parameter vector $\boldsymbol{\beta} \in \mathbb{R}^d$ and classify using the following model:

$$f_{\boldsymbol{\beta}}(\mathbf{x}) = \begin{cases} 1 \text{ if} & \langle \boldsymbol{\beta}, \mathbf{x} \rangle \leq 1 \\ 2 \text{ if} & 2 < \langle \boldsymbol{\beta}, \mathbf{x} \rangle \leq 3 \\ 3 \text{ if} & 3 < \langle \boldsymbol{\beta}, \mathbf{x} \rangle \leq 4 \\ \vdots & \\ q-1 \text{ if} & q-2 < \langle \boldsymbol{\beta}, \mathbf{x} \rangle \leq q-1 \\ q \text{ if} & q-1 < \langle \boldsymbol{\beta}, \mathbf{x} \rangle \end{cases}$$

Describe **one potential issue** and **one potential benefit** of your coworker's method over the approaches mentioned in (a). There is no one "right" answer here.

(c) For the two datasets D_1 and D_2 below, indicate which of the three approaches (one-vs-one, one-vs-all, or your coworkers approach) would lead to an accurate solution to the multiclass classification problem. No explanation is required, but having one might help you earn partial credit.

3. Polynomial transformation

We are given data with just one predictor variable and one target: $(x_1, y_1), \ldots, (x_n, y_n)$, with the goal of fitting a degree two polynomial model using unregularized multiple linear regression with data transformation. The goal is to find the best coefficients $\beta_0, \beta_1, \beta_2$ for predicting y as $\beta_0 + \beta_1 x + \beta_2 x^2$.

Consider the following three transformed data matrices:

$$X_1 = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}, X_2 = \begin{bmatrix} 1 & x_1^2 - x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n^2 - x_n & x_n^2 \end{bmatrix}, \text{ and } X_3 = \begin{bmatrix} 1 & 2x_1^2 - x_1 & 2x_1 - 4x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & 2x_n^2 - x_n & 2x_n - 4x_n^2 \end{bmatrix}$$

Which of the above matrices can be used to solve this problem? In other words, if we train a multiple linear regression problem with X_i can we obtain an optimal degree two polynomial fit for $y_1, \ldots, y_n$. Justify your answer in words, or with equations.

4. Transformed linear models

Write each of the following models as transformed linear models. That is, find a parameter vector $\boldsymbol{\beta}$ in terms of the given parameters a_i and data transformation $\phi(\mathbf{x})$ such that $y = \langle \boldsymbol{\beta}, \phi(\mathbf{x}) \rangle$. Also, show how to recover the original parameters a_i from the parameters β_j:

1. **Example**: $y = a_1 x_1^2 + a_2 \log(a_3 x_2)$. Solution: Notice that $y = a_1 x_1^2 + a_2 \log(x_2) + a_2 \log(a_3)$. Let $\phi([x_1, x_2]) = [x_1^2, \log(x_2), 1]$. Set $a_1 = \beta_1$, $a_2 = \beta_2$, $a_3 = e^{\beta_3/a_2}$.
2. $y = \begin{cases} a_1 + a_2 x & \text{if } x < 1 \\ a_3 + a_4 x & \text{if } x \geq 1 \end{cases}$
3. $y = (1 + a_1 x_1) e^{-x_2 + a_2}$.
4. $y = (a_1 x_1 + a_2 x_2) e^{-x_1 - x_2}$.

5. Estimating a range

Suppose you are given data $x_1, \ldots, x_n$, where each x_i is a single real value. Suppose that the data is distributed independently and **uniformly at random** between $-w$ and w for some real value w, Your goal is to find the maximum likelihood estimate of w based on your data.

1. Write an expression for the likelihood function of the data $\{x_1, \ldots, x_n\}$ given the parameter w.
2. What is the Maximum Likelihood Estimate (MLE) for w based on $\{x_1, \ldots, x_n\}$?