CAP 6610 Machine Learning, Spring 2025

Midterm 2 Solution

Name: ____________________________ **UFID:** ____________________________

This is a 2-hour in-class midterm exam. You may not use any books or notes, but a double-sided cheat sheet is allowed. You will write your exam answers directly in this exam. You should use the attached scratch paper to do your rough work. Feel free to tear them away when submitting.

Problem	Score
1	/20
2	/20
3	/20
4	/20
5	/20

1. (20%) For each of the following statements, determine whether they are true or false. Circle the correct response for each statement below. You do not need to justify your responses.

   (a) Logistic regression is used for classification, not regression.

   $\boxed{\text{True}}$ False

   (b) Gaussian discriminant analysis is a discriminative classifier.

   True $\boxed{\text{False}}$

   (c) The function $1/x$ with domain $\{x \in \mathbb{R} \mid x \neq 0\}$ is a convex function.

   True $\boxed{\text{False}}$

   (d) For a convex and differentiable function, the first-order Taylor approximation at any point is always a global lowerbound.

   $\boxed{\text{True}}$ False

   (e) The sigmoid function $1/(1 + e^{-x})$ is log-concave.

   $\boxed{\text{True}}$ False

2. (20%) *Elementary properties of the quadratic regularized logistic classification.* Consider

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \boldsymbol{\phi}_i^\top \boldsymbol{\theta})) + \lambda \|\boldsymbol{\theta}\|^2, \tag{1}$$

for $y_i = \pm 1$. Circle the correct response for each statement below. You do not need to justify your responses.

   (a) Problem (1) has a unique globally optimal solution.

   $\boxed{\text{True}}$ False

   (b) Let $\boldsymbol{\theta}^\star$ be an optimal solution for (1), $\boldsymbol{\theta}^\star$ is sparse (has many zero entries).

   True $\boxed{\text{False}}$

   (c) If the training data is linearly separable, then some coefficients θ_j might become infinite if $\lambda = 0$.

   $\boxed{\text{True}}$ False

   (d) At optimum, the empirical risk always increases as we increase λ.

   $\boxed{\text{True}}$ False

   (e) On a test set, the prediction accuracy always increases as we increase λ.

   True $\boxed{\text{False}}$

3. (20%) Let x be a real-valued random variable with sample space $\{a_1, \dots, a_k\}$ where $a_1 \le a_2 \le \cdots \le a_k$. This can be viewed as a categorical random variable with each category assigned a real value. Let $\Pr[x = a_i] = p_i$, then the vector $\boldsymbol{p}$ satisfies $\boldsymbol{p} \ge 0$ and $\boldsymbol{1}^\top \boldsymbol{p} = 1$, i.e., it lies in the probability simplex Δ. Each of the following is a function of $\boldsymbol{p}$; as an example, $\mathrm{E}[x] = a_1 p_1 + a_2 p_2 + \cdots + a_k p_k = \boldsymbol{a}^\top \boldsymbol{p}$, which is a linear function in $\boldsymbol{p}$. Determine if each of the following function is convex, concave, or neither.

(a) $\mathrm{E}[x]$

Convex | Concave | $\boxed{\text{Both}}$ | Neither

(b) $\Pr[x > \alpha]$

Convex | Concave | $\boxed{\text{Both}}$ | Neither

(c) $\Pr[\alpha < x < \beta]$

Convex | Concave | $\boxed{\text{Both}}$ | Neither

(d) $-\sum_{i=1}^{k} p_i \log p_i$, the entropy of this distribution

Convex | $\boxed{\text{Concave}}$ | Both | Neither

(e) $\mathrm{var}(x)$

Convex | $\boxed{\text{Concave}}$ | Both | Neither

4. (20%) *Fitting with censored data.* In some experiments there are two kinds of measurements or data available: The usual ones, in which you get a number (say), and censored data, in which you don't get the specific number, but are told something about it, such as a lower bound. A classic example is a study of lifetimes of a set of subjects (say, laboratory mice). For those who have died by the end of data collection, we get the lifetime. For those who have not died by the end of data collection, we do not have the lifetime, but we do have a lower bound, i.e., the length of the study. These are the censored data values.

   We wish to fit a set of data points $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, with $\boldsymbol{x}_i \in \mathbb{R}^m$ and $y \in \mathbb{R}$, using a linear model of the form $y \approx \boldsymbol{x}^\top \boldsymbol{w}$. The vector $\boldsymbol{w} \in \mathbb{R}^m$ is the model parameter, which we wish to choose. We will use a least-squares criterion, i.e., choose $\boldsymbol{w}$ to minimize

$$\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{w})^2.$$

   Here is the tricky part: some of the values of y_i are censored; for these entries, we have only a (given) lower bound. We will re-order the data so that $y_1, \ldots, y_p$ are given (i.e., uncensored), while $y_{p+1}, \ldots, y_n$ are all censored, i.e., unknown, but larger than D, a given number. All the values of $\boldsymbol{x}_i$ are known.

   Formulate an optimization problem to find $\boldsymbol{w}$ (the model parameter) and $y_{p+1}, \ldots, y_n$ (the censored data values). Be sure to clarify the three essential components of an optimization problem: the loss function to be minimized, the optimization variables, and constraints (if any). Is it a convex problem?

Solution.

$$\begin{aligned} \underset{\boldsymbol{w}, y_{p+1}, \ldots, y_n}{\text{minimize}} \quad & \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{w})^2 \\ \text{subject to} \quad & y_i \geq D, i = p+1, \ldots, n. \end{aligned}$$

It is a convex optimization problem. The loss function is convex w.r.t. $y_{p+1}, \ldots, y_n$ and $\boldsymbol{w}$. The constraints are linear.

5. (20%) *Poisson regression.* Consider a regression problem with training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$. Again our predictor takes the form $\widehat{y} = \boldsymbol{\phi}^\top \boldsymbol{\theta}$, where $\boldsymbol{\phi}$ is a feature vector obtained from $\boldsymbol{x}$, and the coefficients $\boldsymbol{\theta}$ is learned from a supervised learning model.

   In many cases the target output y is always a positive integer, so it makes sense to assume that it follows a Poisson distribution. Specifically, we assume that the conditional probability $p(y|\boldsymbol{x})$ follows a Poisson distribution with parameter $\lambda = e^{\boldsymbol{\phi}^\top \boldsymbol{\theta}}$. Recall that the Poisson pmf with parameter λ is

$$p(y) = \frac{\lambda^y e^{-\lambda}}{y!}.$$

   Formulate an optimization problem to learn $\boldsymbol{\theta}$. Is it a convex optimization problem?
   Hint. We will maximize the log-likelihood of the training data, assuming the samples are independent.

   Solution. According to this model,

$$p(y|\boldsymbol{x}) = \frac{e^{y\boldsymbol{\phi}^T\boldsymbol{\theta}} e^{-e^{\boldsymbol{\phi}^T\boldsymbol{\theta}}}}{y!}.$$

   Its log-likelihood of the entire training data is

$$\sum_{i=1}^{n} (y_i \boldsymbol{\phi}_i^T \boldsymbol{\theta} - e^{\boldsymbol{\phi}_i^T \boldsymbol{\theta}} - \log(y_i!)).$$

   This leads to the following optimization problem by maximizing the log-likelihood (notice that the last term does not depend on $\boldsymbol{\theta}$ and thus irrelevant)

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \sum_{i=1}^{n} (e^{\boldsymbol{\phi}_i^T \boldsymbol{\theta}} - y_i \boldsymbol{\phi}_i^T \boldsymbol{\theta}).$$

   This is a convex optimization problem because e^x is convex in x and the objective function is a summation of the affine compositions of this function plus a linear term.