CSE 512: Midterm Oct. 23

Instructions:

- You have 90 continuous minutes to complete the exam, self-timed.
- You are allowed 1 page (front and back) cheat sheet. The cheat sheet must be scanned (or photographed with high resolution) and submitted along with the exam solutions. The time begins when you flip this page.
- You are also allowed a simple calculator. (You may use MATLAB or Python if you agree to only use the simple calculator functions.)
- You may print the exam and write your solutions or use lyx/latex. If you need extra sheets of paper, please label them carefully as to which question they are answering. Make your final answer clear.
- If you choose to handwrite your solutions, you must make sure that the digital scan / photograph is of high enough quality that we can see everything clearly. Anything we can't read, we will not grade.
- You may not discuss any problem with any other student while the exam submission portal is still open. You may not look for answers on the internet or in any notes outside of your cheatsheet.

Name: ____________________________

Student ID: ____________________________

Time began: ____________________________

Time ended: ____________________________

Scoring	
Q 1	________________ / 20
Q 2	________________ / 20
Q 3	________________ / 20
Q 4	________________ / 20
Q 5	________________ / 20
Total	________________ / 100

1. **Concepts.** You do not need to provide a justification for full credit, but we will read it for partial credit.

   (a) **True** or **False**. Unlike a generative classifier, a discriminative classifier first builds a full probabilistic model of the data and label pair, then uses it to return the most likely label of a new point.

   Ans. (4 pts) False. This describes a generative classifier.

   (b) **True** or **False**. Linear regression can only be used to fit straight lines to data.

   Ans. (4 pts) False. Generalized linear regression allows for linear fits of nonlinear bases functions, which are not necessarily straight lines. (Give 3 points if they mark True, but correctly mentioned that generalized linear regression would allow it.)

   (c) **True** or **False**. A common criticism of the maximum likelihood estimator, as compared to the maximum a posteriori estimator, is that it does not account for model parameter randomness.

   Ans. (4 pts) True.

   (d) **True** or **False**. A Bayes classifier is optimal in the worst-case sense.

   Ans. (4 pts) False. The Bayes classifier is optimal in the average sense.

   (e) **True** or **False**. Logistic regression is about fitting a linear model to data, for classification.

   Ans. (4 pts) True. The training loss function is not linear, but the final model offered is linear.

2. **Models for classification** You went out hiking and something bit you. You are trying to decide whether or not to go to the hospital. Below is a dataset which you can use to build a classifier.

Hospital needed (y)	yes (+1)	yes (+1)	yes (+1)	no (-1)	no (-1)	no (-1)
Size of bite ($x[1]$)	5 cm	1 cm	100 mm	10 mm	5 mm	1 mm
Personal allergies ($x[2]$)	yes (+1)	yes (+1)	no (-1)	no (-1)	no (-1)	no (-1)
Hair length ($x[3]$)	1 inch	5 feet	1 mm	1 foot	5.5 feet	bald (0)

(a) Looking at the features *one at a time*, would you propose a linear, generalized linear, or no model to predict y given $x[k]$? Pick the simplest model. Justify each answer.

Feature $x[1]$: linear, generalized linear, or none?
Ans. (5 pts) Linear or generalized linear will work here, although linear is simpler.
Feature $x[2]$: linear, generalized linear, or none?
Ans. (5 pts) Linear model. Matching simply $y = x[2]$ already makes some progress.
Feature $x[3]$: linear, generalized linear, or none?
Ans. (5 pts) No model proposed, since it looks like just noise.

(b) Suppose I decided not to listen to your advice, and trained a logistic regression model using the data as is, no preprocessing. I see the following values for θ^*:

$$\theta^* = \begin{bmatrix} -1.2 \\ 5. \\ -0.2 \end{bmatrix}$$

Based on these readings, how would you rank the importance of each feature, under the assumption of linear modeling?

Ans. (5 pts) In order of decreasing importance,

i. Personal allergies
ii. Size of bite
iii. Hair length

3. **Gradient descent** Consider the following loss function

$$f(\theta) = \frac{1}{m}\sum_{i=1}^{m} \exp(-y_i x_i^T \theta)$$

where $\theta \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$.

(a) Write out the gradient and Hessian of f, giving dimensions.

Ans. (5 pts) Writing $z_i = y_i x_i$ and constructing $Z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix}$, and writing $d_i(\theta) = \exp(-z_i^T\theta)$ for $i = 1, ..., \theta$ with $d = [d_1, d_2, ..., d_m]^T$, we can write the gradient and Hessian of f as

$$\nabla f(\theta) = -Z^T d, \qquad \nabla^2 f(\theta) = Z^T D Z$$

where $D = \mathbf{diag}(d)$.

(b) Show that f is convex.

Ans. (5 pts) f is convex if the Hessian is positive semidefinite. To show that this is true, pick any u. Then

$$u^T \nabla^2 f(\theta) u = \|v\|_2^2, \qquad v = D^{1/2} Z u, \qquad D_{kk}^{1/2} = \sqrt{d_k}$$

which is always nonnegative. Therefore $\nabla^2 f(\theta)$ is positive semidefinite for all θ, and thus f is convex.

Alternative student-inspired solutions: Another way to show that $\nabla^2 f(\theta)$ is positive semidefinite is to use properties of the set of PSD matrices. First, the matrix $z_i z_i^T$ is PSD. Then for any matrix that is PSD, multiplying it by a positive scalar keeps it PSD. Finally, the sum of PSD matrices is PSD. Therefore, since $d_i \geq 0$ for all i, then the matrix $\nabla^2 f(\theta) = \sum_{i=1}^m d_i z_i z_i^T$ is PSD.

A third way is to use properties of convex functions. The scalar function $\exp(-s)$ is convex w.r.t. s. We can also write $f_i(\theta) = \exp(-z_i^T\theta)$ as a convex composition on an affine function, which we also know to be convex. Finally, a sum of convex functions is still convex.

(c) (Hard) Show that f is not L-smooth. (You will need to assume that at least one $x_i \neq 0$.)

Ans. (5 pts) To see that f cannot be L-smooth, we need to show that we can always construct a θ where the largest eigenvalue of the Hessian is unbounded above (can be arbitrarily big). We do this by first assuming that there is some $x_j \neq 0$, and we pick u and θ such that $u^T x_j \neq 0$ and $\theta^T x_j \neq 0$. Then

$$u^T \nabla^2 f(\theta) u = \sum_{i=1}^{m} (z_i u)^2 \cdot \exp(-y_i x_i^T \theta) \geq (z_j^T u)^2 \cdot \exp(-y_j x_j^T \theta)$$

where the last inequality comes from the fact that each term in the sum is positive, so any one term is smaller than the sum.

Assume that f is L-smooth. Then pick $\hat{\theta} = -\frac{1}{y_j x_j^T \theta} \log\left(\frac{2L}{(z_j^T u)^2}\right)\theta$. Then following the same logic,

$$u^T \nabla^2 f(\hat{\theta}) u \geq (z_j^T u)^2 \cdot \exp(-y_j x_j^T \hat{\theta}) = 2L > L.$$

Thus, we found a $\hat{\theta}$ where the largest eigenvalue of it must be larger than L; therefore f cannot be L smooth.

(d) (Really hard) Suppose we further impose that $\|\theta\|_2 \leq 1$. Suppose also that for all i, $x_i^T x_i = 1$. Furthermore, we assume that the matrix with rows $y_i x_i^T$ has full column rank, and smallest singular value $\sigma_{\min}$ and largest singular value $\sigma_{\max}$. Now f is L-smooth and μ-smooth. Find the values of L and μ.

Ans. (5 pts) Using the hint,

$$\begin{aligned}
\max_{u:\|u\|_2=1} u^T\nabla^2 f(\theta)u &= \max_{u:\|u\|_2=1}\sum_{i=1}^m (z_i u)^2 \cdot \exp(-y_i x_i^T\theta) \\
&\leq \max_{u:\|u\|_2=1}\sum_{i=1}^m (z_i u)^2 \cdot \exp(\underbrace{\|y_i x_i^T\|_2\|\theta\|_2}_{\leq 1}) \\
&= \exp(1)\underbrace{\max_{u:\|u\|_2=1}\sum_{i=1}^m (z_i u)^2}_{\sigma_{\max}^2} \\
&= \exp(1)\cdot\sigma_{\max}^2.
\end{aligned}$$

Therefore, we can take $L = \exp(1)\cdot\sigma_{\max}^2$.

$$\begin{aligned}
\min_{u:\|u\|_2=1} u^T\nabla^2 f(\theta)u &= \min_{u:\|u\|_2=1}\sum_{i=1}^m (z_i u)^2 \cdot \exp(-y_i x_i^T\theta) \\
&\geq \min_{u:\|u\|_2=1}\sum_{i=1}^m (z_i u)^2 \cdot \exp(-\underbrace{\|y_i x_i^T\|_2\|\theta\|_2}_{\leq 1}) \\
&= \exp(-1)\underbrace{\min_{u:\|u\|_2=1}\sum_{i=1}^m (z_i u)^2}_{\sigma_{\min}^2} \\
&= \exp(-1)\cdot\sigma_{\min}^2.
\end{aligned}$$

Therefore, we can take $\mu = \exp(-1)\cdot\sigma_{\min}^2$.

4. **Support vector machines** A common depiction of the SVM problem formulation is

$$\begin{array}{ll} \underset{\theta\in\mathbb{R}^n, s\in\mathbb{R}^m}{\text{minimize}} & \|\theta\|_2^2 + \lambda \sum_{i=1}^{m} \max\{0, s_i\} \\ \text{subject to} & y_i x_i^T \theta = 1 - s_i,\ s = 1, ..., m \end{array} \tag{1}$$

Suppose I solve (1) and I receive the following optimal solutions for θ^*, s^*

$$\theta^* = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}, \qquad s^* = \begin{bmatrix} 1.1 \\ 0 \\ -0.5 \\ 0 \\ -3.90 \\ 0 \end{bmatrix}$$

(a) What is the margin of the classifiers I received?

Ans. (5 pts) Margin $= \frac{1}{\|\theta^*\|_2} = \frac{1}{\sqrt{1+9+25}} = \frac{1}{\sqrt{35}} \approx 0.169$

(b) Which points $x_1, x_2, x_3, x_4, x_5, x_6$ are misclassified?

Ans. (5 pts) These are exactly the points where $s_i^* > 0$, namely x_1.

(c) Would you expect my number of misclassified points to increase or decrease if I increase/decrease λ?

Ans. (5 pts) If I increase my penalty on the misclassified slack variables, I would be stricter and de-incentivizing points to be misclassified. Therefore increasing λ should lead to a decreases in number of misclassified points, and decreasing λ to an increase in number of misclassified points.

(d) A cat walks across my keyboard and accidentally deletes θ^* and s^*, but preserves the sparsity pattern (position of nonzero entries) of s^*. Build a linear system that, after solving it, would reveal θ^*.

Ans. (5 pts) Knowing the sparsity pattern, I can just pull out the samples where $s_i^* = 0$, e.g. the points that exactly define the margin. This corresponds to x_2, x_4, and x_6. The linear system can then be set up as

$$\begin{bmatrix} y_2 x_2^T \\ y_4 x_4^T \\ y_6 x_6^T \end{bmatrix} \theta = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

5. **Election season.** I am a professional polling analyst, and my job is to predict who the people of Suffolk county will vote for, in the year 3035. The options are

   - vote for Candidate Martian,
   - vote for Candidate Astroid Belt.

   Suppose that there is no self-reporting bias; everyone I call answers my questions honestly. And, everyone is going to vote; no abstaining. And, assume that no one is conspiring with each other; each vote is given independently.

   I have now randomly polled 500 people, who have given me their voting reports, and 355 of the people I polled claimed they will vote for Candidate Astroid Belt, with 145 of people claiming favor for Candidate Martian.

   (a) Give the maximum likelihood estimate of the true percent of people who will vote for Candidate Martian.
   Ans. (10 pts) $\hat{\theta}_{\text{MLE}} = \frac{145}{500} = 0.29$.

   (b) How certain am I that this estimator is correct, within 5% error margin? That is, if my maximum likelihood estimator is $\hat{X}$, what is the probability that $\hat{X} - 5\% \leq \mathbb{E}[X] \leq \hat{X} + 5\%$?
   As a reminder, here is Hoeffding's inequality

$$\mathbf{Pr}\left(\frac{1}{m}\sum_{i=1}^{m} x_i - \mathbb{E}[X] \geq \epsilon\right) \leq \exp(-2m\epsilon^2).$$

   Ans. (10 pts) Rearranging Hoeffding's inequality and using a union bound, we have

$$\mathbf{Pr}\left(\hat{X} - 5\% \leq \mathbb{E}[X] \leq \hat{X} + 5\%\right) \geq 1 - 2\mathbf{Pr}\left(\frac{1}{m}\sum_{i=1}^{m} x_i - \mathbb{E}[X] \geq \epsilon\right) \geq 1 - 2\exp(-2m\epsilon^2) \approx 0.836$$

   after plugging in $m = 500$ and $\epsilon = 0.05$.