CS 189 Fall 2016 Introduction to Machine Learning Midterm

- Do not open the exam before you are instructed to do so.
- The exam is closed book, closed notes except your one-page cheat sheet.
- Usage of electronic devices is forbidden. If we see you using an electronic device (phone, laptop, etc.) you will get a zero.
- You have 1 hour and 20 minutes.
- Write your initials at the top right of each page (e.g., write "BR" if you are Ben Recht).
- Mark your answers on the exam itself in the space provided. Do not attach any extra sheets.

First name	
Last name	
SID	
First and last name of student to your left	
First and last name of student to your right	

Q1. [25 pts] Clip Loss

Let $S = \{(x_1, y_1), \ldots (x_n, y_n)\}$ be a set of n points sampled i.i.d. from a distribution $\mathcal{D}$. This is the training set with $x_i \in \mathbb{R}^d$ being the features and $y_i \in \{-1, 1\}$ being the labels. Define the *clip loss* of a linear classifier $w \in \mathbb{R}^d$ as

$$\text{loss}(w^\mathsf{T} x, y) = \text{clip}(yw^T x)$$

Where clip is the function

$$\text{clip}(z) = \begin{cases} 1 & \text{if } z < 0 \\ 0 & \text{if } z \geq 1 \\ 1 - z & \text{otherwise.} \end{cases}$$

For any d-dimensional vector w, define the *risk* of w as

$$R[w] = \mathbb{E}_{\mathcal{D}}[\text{loss}(w^\mathsf{T} x, y)]\,,$$

and the *empirical risk* of w as

$$R_S[w] = \frac{1}{n} \sum_{i=1}^{n} \text{loss}(w^\mathsf{T} x_i, y_i)\,.$$

(a) [5 pts] Is the function clip convex? If you would like, you can justify your answer by drawing a picture. It is not convex. Drawing the function shows that the line from $(-1, 1)$ to $(1, 0)$ lies below the graph of the clip function.

(b) [5 pts] Show that if $R_S[w] = 0$ and $\|w\|_2^2 < 1$, then the margin of the hyperplane defined by w is greater than 1. The margin of the hyperplane is defined as

$$\frac{\min_{1 \leq i \leq n}(y_i(w^T x))}{\|w\|_2^2}$$

If $R_S[w] = 0$, then the numerator is greater than or equal to 1 for all i. Moreover, if $\|w\|_2^2 < 1$, the denominator is less than 1. Hence, the margin is greater than 1.

(c) [5 pts] Prove that $\mathbb{E}_S[R_S[w]] = R[w]$.

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\text{loss}(w^\mathsf{T}x_i, y_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\text{loss}(w^\mathsf{T}x_i, y_i)\right] = \frac{1}{n}\sum_{i=1}^{n}R[w] = R[w]$$

(d) [5 pts] Prove that $\text{Var}(R_S[w]) \leq \frac{1}{n}$.

$$\begin{aligned}
\text{Var}(R_S[w]) &= \mathbb{E}\left[(R_S[w] - R[w])^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}(\text{loss}(w^\mathsf{T}x_i, y_i) - R[w])(\text{loss}(w^\mathsf{T}x_j, y_j) - R[w]) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(\text{loss}(w^\mathsf{T}x_i, y_i) - R[w])^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[(\text{loss}(w^\mathsf{T}x, y) - R[w])^2\right] \\
&\leq \frac{1}{n}
\end{aligned}$$

Here, the first line is the definition of variance, the second line expands the square, the third line follows because (x_i, y_i) and (x_j, y_j) are independent. The fourth line follows because the (x_i, y_i) are identically distributed. The last line follows because the clip loss is nonnegative and bounded above by 1.

Alternate proof of first 4 steps:

$$\begin{aligned}
\text{Var}(R_S[w]) &= \text{Var}(\frac{1}{n}\sum_{i=1}^{n}\text{loss}(w^\mathsf{T}x_i, y_i)) \\
&= \frac{1}{n^2}\text{Var}(\sum_{i=1}^{n}\text{loss}(w^\mathsf{T}x_i, y_i)) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(\text{loss}(w^\mathsf{T}x_i, y_i), \text{ by i.i.d} \\
&= \frac{1}{n}\text{Var}(\text{loss}(w^\mathsf{T}x, y))
\end{aligned}$$

(e) [5 pts] Is it possible to have a w such that $R_S[w] = 0$, but $R[w] > 0$? Justify your answer. Yes. Consider the case when $n = 1$. Then it is possible to classify the single data point correctly while classifying all of the opposite class incorrectly.

Q2. [25 pts] Regularization

We consider here a discriminative approach for solving the classification problem illustrated in Figure 1.

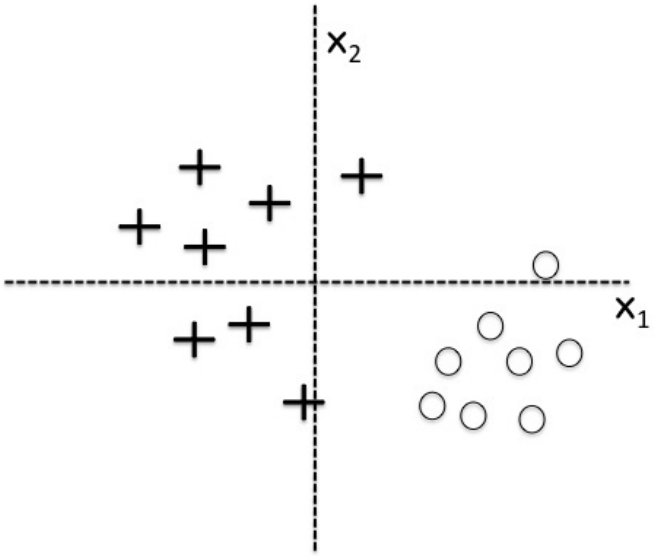

Figure 1: The two-dimensional labeled training set, where '+' corresponds to class $y = 1$ and 'O' corresponds to class $y = 0$.

Suppose we attempt to solve the binary classification task depicted in Figure 1 with the simple linear logistic regression model

$$P(y = 1|x, \mathbf{w}) = g(w_0 + w_1 x_1 + w_2 x_2) = \frac{1}{1 + \exp(-w_0 - w_1 x_1 - w_2 x_2)}$$

Notice that training data can be separated with *zero* training error with a linear separator.

Consider training regularized logistic regression model where we try to maximize

$$\sum_{i=1}^{n} \log P(y_i|x_i, w_0, w_1, w_2) - Cw_j^2$$

for very large C. The regularization penalties used in penalized conditional log-likelihood estimation are $-Cw_j^2$ where $j \in \{0, 1, 2\}$. In other words, only one of the parameters is regularized in each case. Given the training data in Figure 1, how does the training error change with regularization of each parameter w_j? State whether the training error increases or stays the same (zero) for each w_j for large C. Provide a brief justification for each of your answers.

(a) [5 pts] By regularizing w_2

Remains the same. When we regularize w_2, the resulting boundary can rely less and less on the value of x_2 and therefore becomes more vertical and training data can be separated with zero training error with a vertical linear separator.

(b) [5 pts] By regularizing w_1

Increases. When we regularize w_1, the resulting boundary can rely less and less on the value of x_1 and therefore becomes more horizontal. For very large C, the training error increases as there is no good linear horizontal separator of the training data.

(c) [5 pts] By regularizing w_0

Increases. When we regularize w_0, the boundary will eventually go through the origin (bias term set to zero). Based on the figure, we can not find a linear boundary through the origin with zero error. The best we can get is one error.

Now suppose we want to regularize *both* w_1 and w_2. This means we want to maximize the penalized log-likelihood

$$\sum_{i=1}^{n} \log P(y_i|x_i, w_0, w_1, w_2) - C(w_1^2 + w_2^2)$$

Consider again the problem in Figure 1 and the same linear logistic regression model $P(y = 1|x, \mathbf{w}) = g(w_0 + w_1 x_1 + w_2 x_2)$.

(d) [5 pts] For very large C, which value(s) do you expect w_0 to take? Explain briefly. (Note that the number of points from each class is the same.) (You can give a range of values for w_0 if you deem necessary). For very large C, both w_1 and w_2 will go to zero. Note that when $w_1 = w_2 = 0$, the log-probability of labels becomes a finite value, which is equal to $n \log(0.5)$, i.e. $w_0 = 0$. In other words, $P(y = 1|x, \mathbf{w}) = P(y = 0|x, \mathbf{w}) = 0.5$. We expect so because the number of elements in each class is the same and so we would like to predict each one with the same probability, and $w_0 = 0$ makes $P(y = 1|x, \mathbf{w}) = 0.5$.

(e) [5 pts] Assume that we obtain more data points from the '+' class that corresponds to $y = 1$ so that the class labels become unbalanced. Again for very large C, with the same regularization for w_1 and w_2 as above, which value(s) do you expect w_0 to take? Explain briefly. (You can give a range of values for w_0 if you deem necessary). For very large C, we argued that both w_1 and w_2 will go to zero. With unbalanced classes where the number of '+' labels are greater than that of 'o' labels, we want to have $P(y = 1|x, \mathbf{w}) > P(y = 0|x, \mathbf{w})$. For that to happen the value of w_0 should be greater than zero which makes $P(y = 1|x, \mathbf{w}) > 0.5$.

Q3. [25 pts] Bias-variance tradeoff in linear regression

Recall the statistical model for linear regression from lecture. Fix a set of points $x_1, x_2, ..., x_n \in \mathbb{R}^d$ and an unknown regressor $\theta_* \in \mathbb{R}^d$. Suppose we observe $y_1, y_2, ..., y_n \in \mathbb{R}$ via the process

$$y_i = x_i^\mathsf{T}\theta_* + \varepsilon_i,$$

where the noise vector $\varepsilon := \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \in \mathbb{R}^n$ satisfies

$$\mathbb{E}\varepsilon = 0, \qquad \mathrm{Cov}(\varepsilon) = \sigma^2 I_n \,.$$

Using the convention from lecture, we write

$$X := \begin{bmatrix} -x_1^\mathsf{T}- \\ -x_2^\mathsf{T}- \\ \vdots \\ -x_n^\mathsf{T}- \end{bmatrix} \in \mathbb{R}^{n\times d}, \qquad Y := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n \,.$$

With this notation, our statistical model is equivalent to

$$Y = X\theta_* + \varepsilon \,.$$

You may assume throughout this problem that the matrix $X^\mathsf{T}X$ is invertible. Recall the two least-squares estimators we studied in lecture

$$\widehat{\theta}_{\mathrm{ols}} = \arg\min_{\theta\in\mathbb{R}^d} \ \frac{1}{2}\|X\theta - Y\|_2^2 \qquad \text{(OLS)}$$
$$\widehat{\theta}_{\mathrm{ridge}} = \arg\min_{\theta\in\mathbb{R}^d} \ \frac{1}{2}\|X\theta - Y\|_2^2 + \frac{\lambda}{2}\|\theta\|_2^2 \qquad \text{(Ridge)} \,.$$

For the Ridge estimator, you can assume that $\lambda > 0$ is known and fixed throughout the problem.

(a) [5 pts] Write down the closed form solutions for $\widehat{\theta}_{\mathrm{ols}}$ and $\widehat{\theta}_{\mathrm{ridge}}$. Simply state the answer, no need to rederive it.

Answer:

$$\widehat{\theta}_{\mathrm{ols}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$$
$$\widehat{\theta}_{\mathrm{ridge}} = (X^\mathsf{T}X + \lambda I_d)^{-1}X^\mathsf{T}Y \,.$$

(b) [5 pts] Let $\widehat{\theta} \in \mathbb{R}^d$ denote any estimator of θ_*. In the context of this problem, an estimator $\widehat{\theta} = \widehat{\theta}(X, Y)$ is any function which takes the data X and a realization of Y, and computes a guess of θ_*.

Define the MSE (mean squared error) of the estimator $\widehat{\theta}$ as

$$\mathrm{MSE}(\widehat{\theta}) := \mathbb{E}\|\widehat{\theta} - \theta_*\|_2^2 \,.$$

Above, the expectation is taken w.r.t. the randomness inherent in ε. Define $\widehat{\mu} := \mathbb{E}\widehat{\theta}$. Show that, as we did in lecture, the MSE decomposes as such

$$\mathrm{MSE}(\widehat{\theta}) = \|\widehat{\mu} - \theta_*\|_2^2 + \mathbf{Tr}(\mathrm{Cov}(\widehat{\theta})) \,.$$

Hint: Expectation and trace commute, so $\mathbb{E}\,\mathbf{Tr}(A) = \mathbf{Tr}(\mathbb{E}A)$ for any square matrix A.

Answer:

$$\begin{aligned}
\mathbb{E}\|\widehat{\theta} - \theta_*\|_2^2 &= \mathbb{E}\|(\widehat{\theta} - \widehat{\mu}) - (\theta_* - \widehat{\mu})\|_2^2 \\
&= \mathbb{E}\|\widehat{\theta} - \widehat{\mu}\|^2 - 2\mathbb{E}\langle \widehat{\theta} - \widehat{\mu}, \theta_* - \widehat{\mu}\rangle + \mathbb{E}\|\theta_* - \widehat{\mu}\|_2^2 \\
&= \mathbb{E}\|\widehat{\theta} - \widehat{\mu}\|^2 + \|\theta_* - \widehat{\mu}\|_2^2 \\
&= \mathbb{E}\,\mathbf{Tr}((\widehat{\theta} - \widehat{\mu})(\widehat{\theta} - \widehat{\mu})^\mathsf{T}) + \|\theta_* - \widehat{\mu}\|_2^2 \\
&= \mathbf{Tr}(\mathbb{E}(\widehat{\theta} - \widehat{\mu})(\widehat{\theta} - \widehat{\mu})^\mathsf{T}) + \|\theta_* - \widehat{\mu}\|_2^2 \\
&= \mathbf{Tr}(\mathrm{Cov}(\widehat{\theta})) + \|\theta_* - \widehat{\mu}\|_2^2 \,.
\end{aligned}$$

(c) [5 pts] Show that

$$\mathbb{E}\widehat{\theta}_{\mathrm{ols}} = \theta_*, \qquad \mathbb{E}\widehat{\theta}_{\mathrm{ridge}} = (X^\mathsf{T}X + \lambda I_d)^{-1}X^\mathsf{T}X\theta_* \,.$$

That is, $\widehat{\theta}_{\mathrm{ols}}$ is an *unbiased* estimator of θ_*, whereas $\widehat{\theta}_{\mathrm{ridge}}$ is a *biased* estimator of θ_*.

Answer: For OLS,

$$\begin{aligned}
\widehat{\theta}_{\mathrm{ols}} &= (X^\mathsf{T}X)^{-1}X^\mathsf{T}Y \\
&= (X^\mathsf{T}X)^{-1}X^\mathsf{T}(X\theta_* + \varepsilon) \\
&= \theta_* + (X^\mathsf{T}X)^{-1}X^\mathsf{T}\varepsilon \,.
\end{aligned}$$

Hence, since $\mathbb{E}\varepsilon = 0$, $\mathbb{E}\widehat{\theta}_{\mathrm{ols}} = \theta_*$.

Similarly,

$$\begin{aligned}
\widehat{\theta}_{\mathrm{ridge}} &= (X^\mathsf{T}X + \lambda I_d)^{-1}X^\mathsf{T}Y \\
&= (X^\mathsf{T}X + \lambda I_d)^{-1}X^\mathsf{T}(X\theta_* + \varepsilon) \\
&= (X^\mathsf{T}X + \lambda I_d)^{-1}X^\mathsf{T}X\theta_* + (X^\mathsf{T}X + \lambda I_d)^{-1}X^\mathsf{T}\varepsilon \,,
\end{aligned}$$

and therefore $\mathbb{E}\widehat{\theta}_{\mathrm{ridge}} = (X^\mathsf{T}X + \lambda I_d)^{-1}X^\mathsf{T}X\theta_*$.

(d) [10 pts] Let $\gamma_1 \geq \gamma_2 \geq ... \geq \gamma_d$ denote the d eigenvalues of the matrix $X^\mathsf{T}X$ arranged in non-increasing order. First, argue that the smallest eigenvalue, γ_d, is positive (i.e. $\gamma_d > 0$). Then, show that

$$\mathbf{Tr}(\mathrm{Cov}(\widehat{\theta}_{\mathrm{ols}})) = \sigma^2 \sum_{i=1}^{d} \frac{1}{\gamma_i}, \qquad \mathbf{Tr}(\mathrm{Cov}(\widehat{\theta}_{\mathrm{ridge}})) = \sigma^2 \sum_{i=1}^{d} \frac{\gamma_i}{(\gamma_i + \lambda)^2} \,.$$

Finally, use these formulas to conclude that

$$\mathbf{Tr}(\mathrm{Cov}(\widehat{\theta}_{\mathrm{ridge}})) < \mathbf{Tr}(\mathrm{Cov}(\widehat{\theta}_{\mathrm{ols}})) \,.$$

NOTE: The inequality above was *incorrectly* stated on the exam. It is fixed in the solutions. Because of this, we are awarding one free point for every student regardless of whether or not they attempted this question. *Hint:* For the Ridge variance, consider writing $X^\mathsf{T}X$ in terms of its eigen-decomposition $U\Sigma U^\mathsf{T}$.

Answer: For OLS, we simply compute

$$\begin{aligned}
\mathbf{Tr}(\mathbb{E}(X^\mathsf{T}X)^{-1}X^\mathsf{T}\varepsilon\varepsilon^\mathsf{T}X(X^\mathsf{T}X)^{-1}) &= \sigma^2 \,\mathbf{Tr}((X^\mathsf{T}X)^{-1}X^\mathsf{T}X(X^\mathsf{T}X)^{-1}) \\
&= \sigma^2 \,\mathbf{Tr}((X^\mathsf{T}X)^{-1}) \\
&= \sigma^2 \sum_{i=1}^{d} \frac{1}{\gamma_i} \,.
\end{aligned}$$

For Ridge, writing $X^\mathsf{T}X = U\Sigma U^\mathsf{T}$, observe that

$$\begin{aligned}
(X^\mathsf{T}X + \lambda I_d)^{-1} &= U(\Sigma + \lambda I_d)^{-1}U^\mathsf{T} \\
(X^\mathsf{T}X + \lambda I_d)^{-1}X^\mathsf{T}X &= U(\Sigma + \lambda I_d)^{-1}\Sigma U^\mathsf{T} \,.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbf{Tr}(\mathbb{E}(X^\mathsf{T}X + \lambda I_d)^{-1}X^\mathsf{T}\varepsilon\varepsilon^\mathsf{T}X(X^\mathsf{T}X + \lambda I_d)^{-1}) &= \sigma^2 \,\mathbf{Tr}((X^\mathsf{T}X + \lambda I_d)^{-1}X^\mathsf{T}X(X^\mathsf{T}X + \lambda I_d)^{-1}) \\
&= \sigma^2 \,\mathbf{Tr}(U(\Sigma + \lambda I_d)^{-1}\Sigma(\Sigma + \lambda I_d)^{-1}U^\mathsf{T}) \\
&= \sigma^2 \,\mathbf{Tr}(\Sigma(\Sigma + \lambda I_d)^{-2}) \\
&= \sigma^2 \sum_{i=1}^{d} \frac{\gamma_i}{(\gamma_i + \lambda)^2} \,.
\end{aligned}$$

The inequality $\mathbf{Tr}(\mathrm{Cov}(\widehat{\theta}_{\mathrm{ridge}})) < \mathbf{Tr}(\mathrm{Cov}(\widehat{\theta}_{\mathrm{ols}}))$ holds because $(\gamma_i + \lambda)^2 > \gamma_i^2$ for all $1 \leq i \leq d$.

Q4. [15 pts] Nonlinear regression

In this problem, we'll come up with a method for estimating the acceleration on an object from noisy measurements of its position. We have noisy observations of the position of an object ($p(t)$) (in 1D) at n points in time:

$$p_v = [p_0, \ldots, p_n]$$
$$t_v = [t_0, \ldots, t_n].$$

p_i is a noisy measurement of $p(t_i)$. We believe the object is undergoing constant acceleration ($a = p''(t)$).

Note that this means that $p(t) = \frac{1}{2}at^2 + v(0)t + p(0)$ where $v(0)$ is the initial velocity and $p(0)$ is the initial position of the object.

(a) [5 pts] Suppose that $v(0) = 0$ and $p(0) = 0$. Write down a least squares problem to estimate a from the noisy measurements p_v.

$$\min_a \|(\frac{1}{2}t_v.^2)a - p_v\|_2^2.$$

(b) [5 pts] Solve the optimization problem and give your answer in terms of p_v and t_v.

$$\hat{a} = \frac{2\sum_i t_i^2 p_i}{\sum_i t_i^4}.$$

(c) [5 pts] Now suppose the initial position and the velocity of the object are unknown; write down (but don't solve) a least squares problem to estimate the initial position and velocity along with the acceleration of the object.

$$A = [1, t_v, \frac{1}{2}t_v.^2], y = p_v.$$