CS 189 Introduction to Machine Learning
Spring 2022 Marvin Zhang

HW1

Due 2/13/22 at 11:59pm

- Homework 1 consists of all written questions.
- We prefer that you typeset your answers using LaTeX or other word processing software. If you haven't yet learned LaTeX, one of the crown jewels of computer science, now is a good time! Neatly handwritten and scanned solutions will also be accepted for the written questions.
- In all of the questions, **show your work**, not just the final answer.

Deliverables:

1. Submit a PDF of your homework to the Gradescope assignment entitled "HW1 Write-Up". **Please start each question on a new page.** If there are graphs, include those graphs in the correct sections. **Do not** put them in an appendix. We need each solution to be self-contained on pages of its own.
   - In your write-up, please state with whom you worked on the homework. This should be on its own page and should be the first page that you submit.
   - In your write-up, please copy the following statement and sign your signature next to it. (Mac Preview and FoxIt PDF Reader, among others, have tools to let you sign a PDF file.) We want to make it *extra* clear so that no one inadvertently cheats. *"I certify that all solutions are entirely in my own words and that I have not looked at another student's solutions. I have given credit to all external sources I consulted."*
   - **Replicate all your code in an appendix**. Begin code for each coding question in a fresh page. Do not put code from multiple questions in the same page. When you upload this PDF on Gradescope, *make sure* that you assign the relevant pages of your code from appendix to correct questions.

1 Administrivia (2 points)

1. Please fill out the Check-In Survey if you haven't already. Please write down the 10 digit alphanumeric code you get after completing the survey.

2 Multivariate Gaussians: A Review (12 points)

(a) (4 points) Consider a two dimensional random variable $Z \in \mathbb{R}^2$. In order for the random variable to be jointly Gaussian, a necessary and sufficient condition is that

- Z_1 and Z_2 are each marginally Gaussian, and
- $Z_1 | Z_2 = z$ is Gaussian, and $Z_2 | Z_1 = z$ is Gaussian.

A second characterization of a jointly Gaussian RV $Z \in \mathbb{R}^2$ is that it can be written as $Z = AX$, where $X \in \mathbb{R}^2$ is a collection of i.i.d. standard normal RVs and $A \in \mathbb{R}^{2\times 2}$ is a matrix.

Note that the probability density function of a multivariate Gaussian RV with mean vector, μ, and covariance matrix, Σ, is:

$$f(\mathbf{z}) = \exp\left(-\frac{1}{2}(\mathbf{z}-\mu)^T\Sigma^{-1}(\mathbf{z}-\mu)\right) / \sqrt{(2\pi)^k|\Sigma|}$$

.

Let X_1 and X_2 be i.i.d. standard normal RVs. Let U denote a binary random variable uniformly distributed on $\{-1, 1\}$, independent of everything else. Use one of the two characterizations given above to determine whether the following RVs are jointly Gaussian, and calculate the covariance matrix (regardless of whether the RVs are jointly Gaussian).

- $Z_1 = X_1$ and $Z_2 = X_2$.
- $Z_1 = X_1$ and $Z_2 = X_1 + X_2$.
- $Z_1 = X_1$ and $Z_2 = -X_1$.
- $Z_1 = X_1$ and $Z_2 = UX_1$.

(b) (2 points)Use the above example to show that two Gaussian random variables can be uncorrelated, but not independent. On the other hand, show that two uncorrelated, jointly Gaussian RVs are independent.

(c) (2 points) With the setup above, let $Z = VX$, where $V \in \mathbb{R}^{2\times 2}$, and $Z, X \in \mathbb{R}^2$. What is the covariance matrix Σ_Z? Is this also true for a RV other than Gaussian?

(d) (2 points) Use the above setup to show that $X_1 + X_2$ and $X_1 - X_2$ are independent. Give another example pair of linear combinations that are independent.

(e) (2 points) Given a jointly Gaussian RV $Z \in \mathbb{R}^2$ with covariance matrix $\Sigma_Z = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}$, how would you derive the distribution of $Z_1 | Z_2 = z$?

Hint: The following identity may be useful

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{b}{c} & 1 \end{bmatrix} \begin{bmatrix} \left(a - \frac{b^2}{c}\right)^{-1} & 0 \\ 0 & \frac{1}{c} \end{bmatrix} \begin{bmatrix} 1 & -\frac{b}{c} \\ 0 & 1 \end{bmatrix}.$$

3 Gaussian Classification (4 points)

Let $f(x \mid C_i) \sim \mathcal{N}(\mu_i, \sigma^2)$ for a two-class, one-dimensional classification problem with classes C_1 and C_2, $P(C_1) = P(C_2) = 1/2$, and $\mu_2 > \mu_1$.

(a) (2 points) Find the Bayes optimal decision boundary and the corresponding Bayes decision rule. The Bayes optimal decision boundary is defined as the point where $P(C_1|x) = P(C_2|x)$. The Bayes decision rule is how we use the decision boundary to classify a general sample point x.

(b) (2 points) The Bayes error is the probability of misclassification,

$$P_e = P((\text{misclassified as } C_1) \mid C_2)\, P(C_2) + P((\text{misclassified as } C_2) \mid C_1)\, P(C_1).$$

Show that the Bayes error associated with this decision rule is

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-z^2/2} dz$$

where $a = \dfrac{\mu_2 - \mu_1}{2\sigma}$.

4 ℓ_1- and ℓ_2-Regularization (10 points)

Consider sample points $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ and associated values $y_1, y_2, \ldots, y_n \in \mathbb{R}$, an $n \times d$ design matrix $X = [X_1 \quad \ldots \quad X_n]^\top$ and an n-vector $y = [y_1 \quad \ldots \quad y_n]^\top$.

For the sake of simplicity, assume (1) that the sample data have been centered (i.e each feature has mean 0) and (2) that the sample data have been whitened, meaning a linear transformation of is applied to the original data matrix so that the resulting features have variance 1 and the features are uncorrelated; i.e., $X^\top X = nI$.

For this question, we will not use a fictitious dimension nor a bias term; our linear regression function will output zero for $x = 0$.

Consider linear least-squares regression with regularization in the ℓ_1-norm, also known as Lasso. The Lasso cost function is

$$J(w) = |Xw - y|^2 + \lambda \|w\|_1$$

where $w \in \mathbb{R}^d$ and $\lambda > 0$ is the regularization parameter. Let $w^* = \arg\min_{w \in \mathbb{R}^d} J(w)$ denote the weights that minimize the cost function.

In the following steps, we will explore the sparsity-promoting property of the ℓ_1-norm and compare this with the ℓ_2-norm.

1. (2 points) We use the notation X_{*i} to denote column i of the design matrix X, which represents the i^{th} feature. Write $J(w)$ in the following form for appropriate functions g and f.

$$J(w) = g(y) + \sum_{i=1}^{d} f(X_{*i}, w_i, y, \lambda)$$

2. (2 points) If $w_i^* > 0$, what is the value of w_i^*?
3. (2 points) If $w_i^* < 0$, what is the value of w_i^*?
4. (2 points) Considering parts 2 and 3, what is the condition for w_i^* to be zero?
5. (2 points) Now consider ridge regression, which uses the ℓ_2 regularization term $\lambda |w|^2$. How does this change the function $f(\cdot)$ from part 1? What is the new condition in which $w_i^* = 0$? How does it differ from the condition you obtained in part 4?

5 Linear Regression, Projections and Pseudoinverses (10 points)

We are given $X \in \mathbb{R}^{n \times d}$ where $n > d$ and $\text{rank}(X) = d$. We are also given a vector $y \in \mathbb{R}^n$. Define the orthogonal projection of y onto range(X) as $P_X(y)$.

(a) (2 points) An orthogonal projection is a linear transformation. Hence, we can define $P_X(y) = Py$ for some projection matrix P. Specifically, given $1 \leq d \leq n$, a matrix $P \in \mathbb{R}^{n \times n}$ is said to be a rank-d orthogonal projection matrix if $\text{rank}(P) = d$, $P = P^\top$ and $P^2 = P$. Prove that P is a rank-d projection matrix if and only if there exists a $U \in \mathbb{R}^{n \times d}$ such that $P = UU^\top$ and $U^\top U = I$

Hint Use the eigendecomposition of P.

(b) (2 points) Prove that if P is a rank d projection matrix, then $\text{tr}(P) = d$.

(c) (2 points) The Singular Value Decomposition theorem states that we can write any matrix X as

$$X = \sum_{i=1}^{\min\{n,d\}} \sigma_i u_i v_i^\top = \sum_{i:\sigma_i > 0} \sigma_i u_i v_i^\top$$

where $\sigma_i \geq 0$, and $\{u_i\}$ and $\{v_i\}$ are an orthonormal. Show that

(a) $\{v_i : \sigma_i > 0\}$ are an orthonormal basis for the row space of of X

(b) Similarly, $\{u_i : \sigma_i > 0\}$ are an orthonormal basis for the columnspace of X
Hint: consider $X^\top$.

(d) (2 points) Prove that if $X \in \mathbb{R}^{n \times d}$ and $\text{rank}(X) = d$, then $X(X^\top X)^{-1}X^\top$ is a rank-d orthogonal projection matrix. What is the corresponding matrix U?

(e) (2 points) Define the Moore-Penrose pseudoinverse to be the matrix:

$$X^\dagger = \sum_{i:\sigma_i > 0} \sigma_i^{-1} v_i u_i^\top,$$

To what operator does the matrix $X^\dagger X$ correspond? What is $X^\dagger X$ if $\text{rank}(X) = d$? If $\text{rank}(X) = d$ and $n = d$?

6 Geometry of Ridge Regression (16 points)

You recently learned ridge regression and how it differs from ordinary least squares. In this question we will explore useful interpretations and reconceptualizations of ridge regression.

(a) (2 points) Recall that ridge regression can be understood as the unconstrained optimization problem

$$\min_{w} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is a design matrix (data), and $\mathbf{y} \in \mathbb{R}^n$ is the target vector of measurement values.

One way to interpret "ridge regression" is as the ordinary least squares for an augmented data set — i.e. adding a bunch of fake data points to our data. Consider the following augmented measurement vector $\hat{\mathbf{y}}$ and data matrix $\hat{\mathbf{X}}$:

$$\hat{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ \mathbf{0}_d \end{bmatrix} \quad \hat{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I}_d \end{bmatrix},$$

where $\mathbf{0}_d$ is the zero vector in $\mathbb{R}^d$ and $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is the identity matrix. **Show that the optimization problem $\min_w \|\hat{\mathbf{y}} - \hat{\mathbf{X}}\mathbf{w}\|_2^2$ has the same minimizer as (1).**

(b) (2 points) Perhaps more surprisingly, one can achieve the same effect as in the previous part by adding fake features to each data point instead of adding fake data points. Let's construct the augmented design matrix in the following way:

$$\hat{\mathbf{X}} = [\mathbf{X}\ \alpha\mathbf{I_n}]$$

i.e. we stack $\mathbf{X}$ with $\alpha\mathbf{I}_n$ horizontally. Here α is a scalar multiplier. Now our problem is underdetermined: the new dimension $d + n$ is larger than the number of points n. Therefore, there are infinitely many values $\boldsymbol{\eta} \in \mathbb{R}^{d+n}$ for which $\hat{\mathbf{X}}\boldsymbol{\eta} = \mathbf{y}$. Consider the following problem:

$$\min_{\boldsymbol{\eta}} \|\boldsymbol{\eta}\|_2^2 \text{ s.t. } \hat{\mathbf{X}}\boldsymbol{\eta} = \mathbf{y}. \tag{2}$$

Find the α that that if $\boldsymbol{\eta}^*$ is the minimizer of (2), then the first d coordinates of $\boldsymbol{\eta}^*$ form the minimizer of (1).

Can you interpret what the final n coordinates of $\boldsymbol{\eta}^*$ represent?

(c) (2 points) One reason why we might want to have small weights $\mathbf{w}$ has to do with the sensitivity of the predictor to its input. Let $\mathbf{x}$ be a d-dimensional list of features corresponding to a new test point. Our predictor is $\mathbf{w}^\top\mathbf{x}$. **What is an upper bound on how much our prediction could change if we added noise $\boldsymbol{\epsilon} \in \mathbb{R}^d$ to a test point's features $\mathbf{x}$?**

(d) (2 points) We know that the solution to ridge regression (1) is given by $\hat{\mathbf{w}}_\mathbf{r} = (\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$. **What happens when $\lambda \to \infty$?** It is for this reason that sometimes ridge regularization is referred to as "shrinkage."

(e) (2 points) Note that in computing $\hat{\mathbf{w}}_{\mathbf{r}}$, we are trying to invert the matrix $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ instead of the matrix $\mathbf{X}^\top\mathbf{X}$. **If $\mathbf{X}^\top\mathbf{X}$ has eigenvalues $\sigma_1^2, \ldots, \sigma_d^2$, what are the eigenvalues of $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$? Comment on why adding the regularizer term $\lambda\mathbf{I}$ can improve the inversion operation numerically. What happens if X is undertetermined?**

(f) (2 points) Another advantage of ridge regression can be seen for under-determined systems. Say we have the data drawn from a $d = 5$ parameter model, but only have $n = 4$ training samples of it, i.e. $\mathbf{X} \in \mathbb{R}^{4\times5}$. Now this is clearly an underdetermined system, since $n < d$. **Show that ridge regression with $\lambda > 0$ results in a unique solution, whereas ordinary least squares can have an infinite number of solutions.**

[Hint: To make this point, it may be helpful to consider $\mathbf{w} = \mathbf{w}_0 + \mathbf{w}^*$ where $\mathbf{w}_0$ is in the null space of $\mathbf{X}$ and $\mathbf{w}^*$ is a solution.]

[Alternative Hint: You might want to consider (2) *as the way to interpret ridge regression.]*

(g) (2 points) For the previous part, **what will the answer be if you take the limit $\lambda \to 0$ for ridge regression?**

[Hint: You might want to consider (2) *as the way to interpret ridge regression.]*

(h) (2 points) Tikhonov regularization is a general term for ridge regression, where the implicit constraint set takes the form of an ellipsoid instead of a ball. In other words, we solve the optimization problem

$$\mathbf{w} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\Gamma\mathbf{w}\|_2^2$$

for some full rank matrix $\Gamma \in \mathbb{R}^{d\times d}$. **Derive a closed form solution to this problem.**