CS 182: Introduction to Machine Learning, Fall 2022
Homework 1

(Due on Monday, Oct. 10 at 11:59pm (CST))

Notice:

- Please submit your assignments via Gradescope. The entry code is G2V63D.
- Please make sure you select your answer to the corresponding question when submitting your assignments.
- Each person has a total of five days to be late without penalty for all the assignments. Each late delivery less than one day will be counted as one day.

1. [20 points] [Probability Theory]

   (a) Prove that the correlation matrix is positive semidefinite. [6 points]

   (b) Prove that if x_m and x_n are data points sampled from the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, then

   $$\mathbb{E}[x_m x_n] = \mu^2 + I_{mn}\sigma^2,$$

   where $I_{mn} = \begin{cases} 1, & m = n \\ 0, & m \neq n \end{cases}$. [7 points]

   (c) Prove

   $$P(C_i \mid A, B) = \frac{P(C_i, B \mid A)}{\sum_{i=1}^{n} P(C_i, B \mid A)},$$

   where $\{C_i\}_{i=1}^{n}$ is a partition of the sample space. [7 points]

2. [20 points] [Probability Theory, Bayesian Decision Theory] Let $\{x_i\}_{i=1}^N$ be a set of random variables normally distributed with mean μ and variance σ^2, where μ is unknown.

   (a) Derive the maximum likelihood estimate μ_{ML}. [5 points]

   (b) Assume $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Derive the maximum a posteriori estimate μ_{MAP}. [10 points]

   (c) Show that the maximum a posteriori estimate tends to the maximum likelihood estimate ($\mu_{MAP} \to \mu_{ML}$) when $N \to \infty$. [5 points]

3. [20 points] [Introduction, Optimization Primer] Given a set of data points $X = \{\mathbf{x}_i\}_{i=1}^N$, its convex hull is defined as

$$C(X) = \left\{ \mathbf{x} \mid \mathbf{x} = \sum_{i=1}^{N} \theta_i \mathbf{x}_i, \theta_i \geq 0, \sum_{i=1}^{N} \theta_i = 1 \right\}.$$

Similarly we have another data set $Y = \{\mathbf{y}_i\}_{i=1}^N$ and its corresponding convex hull $C(Y)$. Show that if convex hulls of two sets of points intersect, these two sets are not linearly separable, and conversely that if they are linearly separable, their convex hulls do not intersect. (Hint: Two sets of points are linearly separable, if there exists a vector $\mathbf{w}$ and a scalar b such that $\forall\ \mathbf{x}_i$, $\mathbf{w}^\mathsf{T}\mathbf{x}_i + b > 0$ and $\forall\ \mathbf{y}_i$, $\mathbf{w}^\mathsf{T}\mathbf{y}_i + b < 0$.)

4. [20 points] [Linear Algebra] Assume that there are n given training examples $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$, where each input data point $\mathbf{x}_i$ has m real valued features. When $m > n$, the linear regression model is equivalent to solving an under-determined system of linear equations $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$. One popular way to estimate $\boldsymbol{\beta}$ is to consider the so-called ridge regression:

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2,$$

for some $\lambda > 0$. This is also known as Tikhonov regularization.

(a) Show that the optimal solution $\boldsymbol{\beta}_\star$ to the above optimization problem is given by:

$$\boldsymbol{\beta}_\star = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y},$$

given that the matrix $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is invertible. [10 points]

(b) Discuss the conditions on the matrix $\mathbf{X}$ and λ so that the matrix $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is guaranteed to be invertible. [10 points]

5. [20 points] [Optimization Primer]

   (a) Prove that if f is a convex function, then $\mathcal{C} = \{\mathbf{x} \mid f(\mathbf{x}) \leq 0\}$ is a convex set. [10 points]

   (b) Prove that if x is a random variable and f is a convex function, then $f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$. [10 points]