Machine Learning (CS 6140)
Homework 1 Solution

Instructor: Ehsan Elhamifar
Due Date: October 12, 2017, 11:45am

1) Probability and Random Variables: State true or false. If true, prove it. If false, either prove or demonstrate by a counter example. Here Ω denotes the sample space and A^c denotes the complement of the event A.

1. For any $A, B \subseteq \Omega$ such that $P(A) > 0$, $P(B^c) > 0$, $P(B|A) + P(A|B^c) = 1$.
solution: False. Notice that the above is equivalent to show $P(A|B^c) = 1 - P(B|A) = P(B^c|A)$. Using the definition of conditional probability, we have $P(A|B^c) = P(A, B^c)/P(B^c)$ and $P(B^c|A) = P(A, B^c)/P(A)$. Thus, as long as $P(A) \neq P(B^c)$, we have $P(A|B^c) \neq P(B^c|A)$, hence $P(B|A) + P(A|B^c) \neq 1$.

2. For any $A, B \subseteq \Omega$ such that $0 < P(B) < 1$, $P(A|B) + P(A|B^c) = 1$.
solution: False. Let $B = A$ and P(A) = 0.1. We have $P(A|A) + P(A|A^c) = P(A) = 0.1 \neq 1$.

3. For any $A, B \subseteq \Omega$, $P(B^c \cup (A \cap B)) + P(A^c \cap B) = 1$.
solution: True. Notice that $B^c \cup (A \cap B) = (B^c \cup A) \cap (B^c \cup B) = (B^c \cup A) \cap \Omega = (B^c \cup A)$. On the other hand, since $(B^c \cup A)^c = B \cap A^c$, we have $P(B^c \cup (A \cap B)) + P(A^c \cap B) = P(B^c \cup A) + P((B^c \cup A)^c) = 1$.

4. Let $\{A_i\}_{i=1}^n$ be mutually independent. Then, $P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$.
solution: False. Let $n = 2$ and $P(A_1), P(A_2) > 0$. By mutual independence assumption, we have $P(A_1 \cap A_2) = P(A_1)P(A_2)$. On the other hand, $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) = P(A_1) + P(A_2) - P(A_1)P(A_2) < P(A_1) + P(A_2)$.

5. Let $\{(A_i, B_i)\}_{i=1}^n$ be mutually independent, i.e., (A_i, B_i) is independent from (A_j, B_j) for every $i \neq j$. Then $P(A_1, \ldots, A_n|B_1, \ldots, B_n) = \prod_{i=1}^n P(A_i|B_i)$.
solution: True. We can write $P(A_1, ..., A_n|B_1, ..., B_n) = \frac{P(A1,...,A_n,B1,...,B_n)}{P(B1,...,B_n)}$. As (A_i, B_i) is independent from (A_j, B_j) for $i \neq j$, $\frac{P(A_1,...,A_n,B_1,...,B_n)}{P(B_1,...,B_n)} = \frac{\prod_{i=1}^n P(A_i,B_i)}{P(B_1,...,B_n)}$. Since (A_i, B_i) is independent from (A_j, B_j), any subset of the two will be independent from each other, hence, B_i is independent from B_j. As a result, $P(B_1, ..., B_n) = \prod_{i=1}^n P(B_i)$. Hence, $P(A_1, ..., A_n|B_1, ..., B_n) = \frac{\prod_{i=1}^n P(A_i,B_i)}{\prod_{i=1}^n P(B_i)} = \prod_{i=1}^n P(A_i|B_i)$.

2) Discrete and Continuous Distributions: Write down the formula of the probability density/mass functions of random variable X.

1. k-dimensional Gaussian distribution (multi-variate Gaussian), $X \sim \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$.

solution:

$$f(X) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$$

2. Laplace distribution with mean μ and variance $2\sigma^2$.
solution:

$$f(X;\mu,\delta) = \frac{1}{2\delta}\exp(-\frac{|x-\mu|}{\delta})$$

3. Bernoulli distribution, $X \sim \text{Bernoulli}(p), 0 < p < 1$.
solution:

$$f(X;p) = p^X(1-p)^{1-X} \text{ where } X \in 0, 1,\ p \in (0,1)$$

4. Multinomial distribution with N trials and L outcomes with probabilities $\theta_1, \ldots, \theta_L$.
solution:

$$f(X;N,p_1,...,p_L) = \begin{cases} \dfrac{n!}{x_1!\cdots x_L!}p_1^{x_1}\cdots p_L^{x_L} & \text{when } \sum_{i=1}^{L} x_i = n \\ 0 & \text{otherwise} \end{cases}$$

5. Dirichlet distribution of order L with parameters $\alpha_1, \ldots, \alpha_L$.
solution:

$$f(X;L,\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})}\prod_{i=1}^{L} x_i^{\alpha_i-1} \text{ where } B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{L}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{L}\alpha_i)},\ \boldsymbol{\alpha} = (\alpha_1,...,\alpha_L)$$

6. Uniform distribution, $X \sim \text{Unif}(a,b), a < b$.
solution:

$$f(X;a,b) = \begin{cases} \dfrac{1}{b-a} & x \in [a,b],\ a < b \\ 0 & \text{otherwise} \end{cases}$$

7. Exponential distribution, $X \sim \text{Exp}(\lambda), \lambda > 0$.
solution:

$$f(X;\lambda) = \begin{cases} \lambda e^{-\lambda X} & x \geq 0,\ \lambda > 0 \\ 0 & x < 0 \end{cases}$$

8. Poisson distribution, $X \sim \text{Poisson}(\lambda), \lambda > 0$.
solution:

$$f(X;\lambda) = \frac{\lambda^X e^{-\lambda}}{X!},\ \lambda > 0$$

3) Positive-Definite Matrices: A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{m\times m}$ is positive-semidefinite if $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} \geq 0$ for every $\boldsymbol{x} \in \mathbb{R}^m$, where $\boldsymbol{x} \neq \boldsymbol{0}$. Equivalently, $\boldsymbol{A}$ is positive-semidefinite if all eigenvalues of $\boldsymbol{A}$ are non-negative. Prove or disprove (by a counter example) that the following matrices are positive-semidefinite.

1. $\boldsymbol{A} = \boldsymbol{B}^\top \boldsymbol{B}$ for an arbitrary $\boldsymbol{B} \in \mathbb{R}^{m \times n}$
solution: True. $\boldsymbol{A}^\top = (\boldsymbol{B}^\top \boldsymbol{B})^\top = \boldsymbol{B}^\top (\boldsymbol{B}^\top)^\top = \boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{A}$, thus $\boldsymbol{A}$is a symmetric matrix. On the other hand, for every $x \in \mathbb{R}^m, \boldsymbol{x} \neq \boldsymbol{0}$, we have $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{B}^\top \boldsymbol{B}\boldsymbol{x} = (\boldsymbol{B}\boldsymbol{x})^\top \boldsymbol{B}\boldsymbol{x} = \|\boldsymbol{B}\boldsymbol{x}\|_2^2 \geq 0$. Thus $\boldsymbol{A}$ is a positive-semidefinite matrix.

2. $\boldsymbol{A} = \begin{bmatrix} 8 & -5 & -3 \\ -5 & 5 & 0 \\ -3 & 0 & 3 \end{bmatrix}$
solution: True. The eigenvalues of $\boldsymbol{A}$, are determined by solving the characteristic equation $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$, where $\boldsymbol{I}$ is the identity matrix.

$$\boldsymbol{A} - \lambda \boldsymbol{I} = \begin{bmatrix} 8-\lambda & -5 & -3 \\ -5 & 5-\lambda & 0 \\ -3 & 0 & 3-\lambda \end{bmatrix}$$

$$\begin{aligned} \det(\boldsymbol{A} - \lambda \boldsymbol{I}) &= (8-\lambda) \begin{vmatrix} 5-\lambda & 0 \\ 0 & 3-\lambda \end{vmatrix} - (-5) \begin{vmatrix} -5 & 0 \\ -3 & 3-\lambda \end{vmatrix} + (-3) \begin{vmatrix} -5 & 5-\lambda \\ -3 & 0 \end{vmatrix} \\ &= -\lambda(\lambda^2 - 16\lambda + 45) = 0 \implies \lambda_1 = 0,\ \lambda_2 = 3.64,\ \lambda_3 = 12.36. \end{aligned}$$

Since all the eigenvalues are non-negative, $\boldsymbol{A}$ is positive-semidefinite.

3. $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{B}^\top + \boldsymbol{B}^\top \boldsymbol{B}$ for an arbitrary $\boldsymbol{B} \in \mathbb{R}^{n \times n}$
solution: False. Assume $\boldsymbol{B} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$, hence, $A = B + B^T + B^T B = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$. Given $\boldsymbol{x} \neq \boldsymbol{0}$, $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = -\|\boldsymbol{x}\|_2^2 < 0$. Thus $\boldsymbol{A}$ is not positive-semidefinite.

4) Convexity of Linear Regression: In the class, we studied several models for linear regression. Let $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ and $\boldsymbol{Y} \in \mathbb{R}^N$ denote matrices of input features and outputs/responses, respectively. Let $\boldsymbol{\theta} \in \mathbb{R}^d$ denote the vector of unknown parameters.

a) Show that the following objective functions for linear regression are convex with respect to $\boldsymbol{\theta}$.

1. Vanilla/Basic regression: $J_1(\boldsymbol{\theta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2$
solution: A function is convex if the Hessian of the function (second order partial derivatives of the function) is positive-semidefinite. Taking the derivative of J_1 with respect to $\boldsymbol{\theta}$ we get

$$\frac{\partial J_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\boldsymbol{X}^\top \boldsymbol{Y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta}.$$

The Hessian of the loss function is

$$\frac{\partial^2 J_1(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = 2\boldsymbol{X}^\top \boldsymbol{X},$$

which is positive semidefinite. Therefore, $J_1(\boldsymbol{\theta})$ is a convex function.

2. Ridge regression: $J_2(\boldsymbol{\theta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2$
solution: Taking the derivative of J_2 with respect to $\boldsymbol{\theta}$ we have

$$\frac{\partial J_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -2\boldsymbol{X}^\top \boldsymbol{Y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} + 2\lambda \boldsymbol{I}_{d\times d}\boldsymbol{\theta}.$$

The Hessian of the function is then

$$\frac{\partial^2 J_2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = 2(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_{d\times d})$$

The Hessian is the sum of the positive semidefinite matrix $\boldsymbol{X}^\top \boldsymbol{X}$ and the positive definite matrix $\lambda \boldsymbol{I}_{d\times d}$, hence is a positive definite matrix. Therefore, $J_2(\boldsymbol{\theta})$ is a convex function.

3. Lasso regression: $J_3(\boldsymbol{\theta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1$
solution: As we saw in part a, $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2$ is a convex function. On the other hand, any vector norm, including $\|\boldsymbol{\theta}\|_1$ is convex. The sum of two convex functions (since $\lambda > 0$) is always convex.

b) What conditions do we need to impose on $\boldsymbol{X}$ and/or $\boldsymbol{Y}$ in each of the above cases, so that the solution for $\boldsymbol{\theta}$ be unique?
solution: For J_1, we need that $\boldsymbol{X}^\top \boldsymbol{X}$ be invertible in order to have a unique solution for $\boldsymbol{\theta}$ (when setting the gradient to zero). This means that $\boldsymbol{X}^\top \boldsymbol{X}$ must be non-singular, i.e., all its eigenvalues be nonzero. Thus, when all eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$ are positive, i.e., when $\boldsymbol{X}^\top \boldsymbol{X}$ is positive definite, the solution can be determined uniquely.
For J_2, we need invertibility of $(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_{d\times d})$. Since all the eigenvalues of $(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_{d\times d})$ are positive, it is invertible. Hence, minimizing J_2 always gives a unique solution for $\boldsymbol{\theta}$.
For J_3, similar to J_1, there may exist multiple solutions that minimize the cost function.

5) Regression using Huber Loss: In the class, we defined the Huber loss as

$$\ell_\delta(e) = \begin{cases} \frac{1}{2}e^2 & \text{if } |e| \le \delta \\ \delta|e| - \frac{1}{2}\delta^2 & \text{if } |e| > \delta \end{cases}$$

Consider the robust regression model

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \ell_\delta(y_i - \boldsymbol{\theta}^\top \boldsymbol{x}_i),$$

where $\boldsymbol{x}_i$ and y_i denote the i-th input sample and output/response, respectively and $\boldsymbol{\theta}$ is the unknown parameter vector.
a) Provide the steps of the batch gradient descent in order to obtain the solution for $\boldsymbol{\theta}$.
b) Provide the steps of the stochastic gradient descent using mini-batches of size 1, i.e., one sample in each mini-batch, in order to obtain the solution for $\boldsymbol{\theta}$.

solution: The gradient of the Huber loss function is

$$\nabla_e \ell_\delta(e) = \begin{cases} -\delta & \text{if } e < -\delta \\ e & \text{if } -\delta \leq e \leq \delta > \delta \\ +\delta & \text{if } e > \delta \end{cases}$$

Notice that the gradient is continuous at $e = -\delta$ and $e = \delta$. Let $e_i \triangleq y_i - \boldsymbol{\theta}^\top \boldsymbol{x}_i$, then $\nabla_\theta l_\delta(y_i - \theta^\top \boldsymbol{x}_i) = \nabla_{e_i} l_\delta(e_i) \nabla_\theta e_i = \nabla_{e_i} l_\delta(e_i)(-\boldsymbol{x}_i)$. Thus, for the cost function $J(\boldsymbol{\theta}) = \sum_{i=1}^N l_\delta(y_i - \boldsymbol{\theta}^T \boldsymbol{x}_i)$, and $\nabla_\theta J(\theta) = -\sum_{i=1}^N \boldsymbol{x}_i \nabla_{e_i} l_\delta(e_i)$.

Batch gradient descent
Step 1: initialize the parameters vector θ randomly and set up the learning rate α.
Step 2: repeat the following until convergence
Step 2.1: calculate the gradient of the cost function
Step 2.2: update the parameter vector, $\boldsymbol{\theta} := \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$.

Stochastic gradient descent
Step 1: initialize the parameters vector θ randomly and set up the learning rate α.
Step 2: repeat the following steps until convergence
Step 2.1: draw a sample i from $\{1, \ldots, N\}$ uniformly at random, denote the sample as $(\boldsymbol{x}_i, y_i)$. Calculate the gradient of $J(\boldsymbol{\theta}; \boldsymbol{x}_i; y_i) = l_\delta(y_i - \boldsymbol{\theta}^\top \boldsymbol{x}_i)$, and update the parameter vector, $\boldsymbol{\theta} := \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}; \boldsymbol{x}_i; y_i)$.

6) PAC Confidence Bounds: In the class, we studied the problem of maximum likelihood estimation of a Bernoulli random variable (taking values in $\{0, 1\}$), where the true probability of being 1 is assumed to be θ^o. We showed that using the maximum likelihood estimation on a dataset with N samples, the ML estimate is given by $\hat{\theta} = \sum_{i=1}^N x_i / N$. Moreover, we showed that

$$P\left(|\hat{\theta} - \theta^o| \geq \epsilon\right) \leq 2e^{-2N\epsilon^2}.$$

Consider the example of flipping a coin N times with the true probability of 'Head' to be θ^o. How many trials (flipping the coin) we need to have in order to be confident that with probability at least 0.95, the estimate of the maximum likelihood for the probability of 'Head' will be within 0.1 distance of the true value?
solution: We can solve for N as follows:

$$\begin{aligned} P(|\hat{\theta}_{MLE} - \theta^*| \geq 0.1) \leq 2e^{-2N(0.1)^2} &\leq (1 - 0.95) \\ \rightarrow 1 - 0.95 &\geq 2e^{-2N*0.1^2} \\ \rightarrow N &\geq \frac{\log \frac{2}{0.05}}{2 * 0.1^2} \approx 185 \end{aligned}$$

Note: There was a factor 2 missing in the exponential term in the original posted question. If you have solved the problem using that (i.e., you obtained $N \geq 369$), it will also be accepted.

7) Probabilistic Regression with Prior on Parameters: Consider the probabilistic model of regression, where $p(y|\boldsymbol{x}, \boldsymbol{\theta})$ is a Normal distribution with mean $\theta^\top \boldsymbol{x}$ and variance σ^2, i.e., $\mathcal{N}(\boldsymbol{\theta}^\top \boldsymbol{x}, \sigma^2)$. We would like to determine $\boldsymbol{\theta}$ using a dataset of N samples $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$. Assume we have prior information about the distribution of θ and our goal is to determine the Maximum A Posteriori (MAP) estimate of $\boldsymbol{\theta}$ using the dataset and the prior information. For each of the following cases, provide the optimization from which we can obtain the MAP solution.

1. $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, 1/\lambda \boldsymbol{I})$, where $\boldsymbol{I}$ denotes the identity matrix.
solution: To compute MAP, we need to compute $p(\boldsymbol{\theta}|\mathcal{D})$. Notice that $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Assuming that the samples are i.i.d., we have $p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^N p(\boldsymbol{x}_i, y_i|\boldsymbol{\theta}) = \prod_{i=1}^N p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta})p(\boldsymbol{x}_i|\boldsymbol{\theta})$ = $\prod_{i=1}^N p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta})p(\boldsymbol{x}_i) \propto \prod_{i=1}^N p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta})$. Here we used the fact that $\boldsymbol{x}_i$'s are independent from $\boldsymbol{\theta}$. Thus, we have

$$p(\mathcal{D}|\boldsymbol{\theta}) \propto \prod_{i=1}^N p(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \boldsymbol{\theta}^\top \boldsymbol{x}_i)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\frac{\sum_{i=1}^N (y_i - \boldsymbol{\theta}^\top \boldsymbol{x}_i)^2}{2\sigma^2}} \propto e^{-\frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2}{2\sigma^2}},$$

where $\boldsymbol{Y} = \begin{bmatrix} y_1 & \cdots & y_N \end{bmatrix}^\top$ and $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_N \end{bmatrix}^\top$. Using the distribution of $\boldsymbol{\theta}$, we get

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto e^{-\frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2}{2\sigma^2}} e^{-\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2} = e^{-\frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2}{2\sigma^2} - \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2}.$$

Thus, to maximize the posterior and find MAP, we need to minimize

$$\min_{\boldsymbol{\theta}} \frac{1}{2\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2.$$

Taking the derivative of the above cost function with respect to $\boldsymbol{\theta}$ and setting it to zero, we get

$$\frac{1}{\sigma^2}\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{Y}) + \lambda\boldsymbol{\theta} = 0 \implies \boldsymbol{\theta} = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{X}^\top\boldsymbol{Y}).$$

2. Each element of $\boldsymbol{\theta}$ has a Laplace distribution with mean 0 and variance $2/\lambda^2$.
solution: Similar to the previous case, we have

$$p(\mathcal{D}|\boldsymbol{\theta}) \propto e^{-\frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2}{2\sigma^2}}.$$

Since each element j of $\boldsymbol{\theta}$, denoted by θ_j, has a Laplace distribution with mean zero and variance $2/\lambda^2$, we have

$$p(\boldsymbol{\theta}) = \prod_j \frac{\lambda}{2} e^{-\lambda|\theta_j|} \propto \prod_j e^{-\lambda\sum_j|\theta_j|} = e^{-\lambda\|\boldsymbol{\theta}\|_1}.$$

Thus, the posterior can be computed as

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto e^{-\frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2}{2\sigma^2}} e^{-\lambda\|\boldsymbol{\theta}\|_1} = e^{-\frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2}{2\sigma^2} - \lambda\|\boldsymbol{\theta}\|_1}.$$

As a result, to maximize the posterior distribution and find MAP, we need to solve

$$\min_{\boldsymbol{\theta}} \frac{1}{2\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1.$$

8) MAP estimation for the Bernoulli with non-conjugate priors: Consider a Bernoulli random variable x with $p(x = 1) = \theta$. In the class, we discussed MAP estimation of the Bernoulli rate parameter θ with the prior $p(\theta) = \text{Beta}(\theta|\alpha, \beta)$. We know that, with this prior, the MAP estimate is given by:

$$\hat{\theta} = \frac{N_1 + \alpha - 1}{N + \alpha + \beta - 2}$$

where N_1 is the number of trails where $x_i = 1$ (e.g., heads in flipping a coin), N_0 is the number of trials where $x_i = 0$ (e.g., tails in flipping a coin) and $N = N_0 + N_1$ is the total number of trials.

1. Now consider the following prior, that believes the coin is fair, or is slightly biased towards heads:

$$p(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.5 \\ 0.5 & \text{if } \theta = 0.6 \\ 0 & \text{otherwise} \end{cases}$$

Derive the MAP estimate under this prior as a function of N_1 and N.
solution: The posterior probability of θ is:

$$\begin{aligned} P(\theta|D) &= \frac{P(D|\theta)P(\theta)}{P(D)} = \frac{(\prod_{i=1}^{N} \theta^{y_n}(1-\theta)^{1-y_n})p(\theta)}{P(D)} \\ &= \frac{\theta^{N_1}(1-\theta)^{N-N_1}p(\theta)}{P(D)} \end{aligned}$$

So we have:

$$p(\theta = 0.5|D) = \frac{0.5^{N_1}0.5^{N-N_1}0.5}{P(D)} = \frac{0.5^{N}0.5}{P(D)}$$

and

$$p(\theta = 0.6|D) = \frac{0.6^{N_1}0.4^{N-N_1}0.5}{P(D)}$$

Notice that the MAP estimate of θ is

$$\hat{\theta} = \text{argmax}_{\theta} p(\theta|D).$$

Hence, when $p(\theta = 0.5|D) > p(\theta = 0.6|D)$, we get $\hat{\theta} = 0.5$, i.e.,

$$\begin{aligned} p(\theta = 0.5|D) &> p(\theta = 0.6|D) \\ \rightarrow 0.5^N &> 0.6^{N_1}0.4^{N-N_1} \\ \rightarrow (\frac{5}{4})^N &> (\frac{3}{2})^{N_1} \\ \rightarrow N\log(\frac{5}{4}) &> N_1\log(\frac{3}{2}). \end{aligned}$$

As a result, when $N > \frac{\log(\frac{3}{2})}{\log(\frac{5}{4})}N_1 = 1.8N_1$, we have $\hat{\theta} = 0.5$; otherwise, we have $\hat{\theta} = 0.6$.

2. Suppose the true parameters is $\theta = 0.61$. Which prior leads to a better estimate when N is small? Which prior leads to a better estimate when N is large?
   solution: As discussed in the class, when N is large, using the beta prior, the solution of MAP and MLE get very close, hence, Beta would be a good prior. For small N, if we set the values of α and β properly, i.e., when the prior is set properly, the MAP solution can be close to the true value of θ.
   For the new prior, if N is large, then N_1/N will be sufficiently close to the true value $\theta = 0.61$. Thus, $N/N_1 \approx 1.64 < 1.8$. Hence, the MAP using the new prior would be 0.6, which will be close to the true value. However, when N is small, depending on the value of N_1 the MAP can be 0.5 or 0.6.

9) Gaussian Naive Bayes: The multivariate normal distribution in k-dimensions, also called the multi-variate Gaussian distribution and denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, is parameterized by a mean vector $\boldsymbol{\mu} \in \mathbb{R}^k$ and a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}$, where $\boldsymbol{\Sigma} \geq \mathbf{0}$ is positive semi-definite.
Consisder a classification problem in which the input feature $\boldsymbol{x} \in \mathbb{R}^k$ are continuous-valued random variables, we can then use the Gaussian Naive Bayes (GNB) model, which models $p(\boldsymbol{x}|y)$ using a multivariate normal distribution. The model is given by

$$y \sim \text{Bernoulli}(\phi)$$

$$\boldsymbol{x}|y = 0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$$
$$\boldsymbol{x}|y = 1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

where we assume that $\boldsymbol{\Sigma}$ is a diagonal matrix, $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_k^2)$. Given a training dataset $\mathcal{D} = \{(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^N, y^N)\}$, write down the likelihood (log-likelihood) and derive MLE estimates for the means $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, covariance $\boldsymbol{\Sigma}$ and the class prior ϕ of the GNB.
solution: Let $p(y = 1) = \phi$ and $p(y = 0) = 1 - \phi$. Let $\mathcal{N}_1$, N_0 be the number of training samples from class 1 and 0, respectively. The parameters of the model to learn are $\boldsymbol{\theta} \triangleq \{\phi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \sigma_1^2, \ldots, \sigma_k^2\}$. Assuming that the samples are i.i.d., we can write the likelihood function as

$$p_{\boldsymbol{\theta}}(\mathcal{D}) = \prod_{i=1}^{N} p_{\boldsymbol{\theta}}(\boldsymbol{x}^i, y^i) = \prod_{i=1}^{N} p_{\boldsymbol{\theta}}(\boldsymbol{x}^i|y^i) p_{\boldsymbol{\theta}}(y^i).$$

Depending on whether y_i is zero or one, we can rewrite the above as

$$\begin{aligned}
p_{\boldsymbol{\theta}}(\mathcal{D}) &= \prod_{i:y^i=1} p_{\boldsymbol{\theta}}(\boldsymbol{x}^i|y^i = 1) \prod_{i:y^i=0} p_{\boldsymbol{\theta}}(\boldsymbol{x}^i|y^i = 0) \prod_{i:y^i=1} p_{\boldsymbol{\theta}}(y^i = 1) \prod_{i:y^i=0} p_{\boldsymbol{\theta}}(y^i = 0) \\
&= \prod_{i:y^i=1} \mathcal{N}(\boldsymbol{x}^i; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \prod_{i:y^i=0} \mathcal{N}(\boldsymbol{x}^i; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}) \prod_{i:y^i=1} \phi \prod_{i:y^i=0} (1 - \phi) \\
&= \prod_{i:y^i=1} \frac{1}{(2\pi)^{k/2}(\det(\boldsymbol{\Sigma}))^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}^i - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^i - \boldsymbol{\mu}_1)} \times \prod_{i:y^i=0} \frac{1}{(2\pi)^{k/2}(\det(\boldsymbol{\Sigma}))^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}^i - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^i - \boldsymbol{\mu}_0)} \\
&\qquad \times \phi^{N_1} \times (1 - \phi)^{N_0}.
\end{aligned}$$

Taking the logarithm from the above, we have

$$\log(p_{\boldsymbol{\theta}}(\mathcal{D})) = -\frac{N}{2}\log(\det(\boldsymbol{\Sigma})) + \sum_{i:y^i=1} -\frac{1}{2}(\boldsymbol{x}^i - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^i - \boldsymbol{\mu}_1) + \sum_{i:y^i=0} -\frac{1}{2}(\boldsymbol{x}^i - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}^i - \boldsymbol{\mu}_0) + N_1 \log\phi + N_0 \log(1-\phi) + \text{const.}$$

Taking the derivative of the above with respect to ϕ and setting it to zero, we obtain

$$\hat{\phi} = \frac{N_1}{N_1 + N_0} = \frac{N_1}{N}.$$

Taking the derivative with respect to $\boldsymbol{\mu}_1$ and setting it to zero, we obtain

$$\sum_{i:y^i=1} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{x}^i) = 0 \implies \hat{\boldsymbol{\mu}}_1 = \frac{1}{N_1}\sum_{i:y^i=1} \boldsymbol{x}^i.$$

Similarly, we can obtain

$$\sum_{i:y^i=0} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{x}^i) = 0 \implies \hat{\boldsymbol{\mu}}_0 = \frac{1}{N_0}\sum_{i:y^i=0} \boldsymbol{x}^i.$$

Since $\boldsymbol{\Sigma}$ is assumed to be diagonal, we have $\det(\boldsymbol{\Sigma}) = \prod_{j=1}^k \sigma_j^2$. Thus, the log-likelihood can be written as

$$\begin{aligned}\log(p_{\boldsymbol{\theta}}(\mathcal{D})) &= -\frac{N}{2}\sum_{j=1}^k \log\sigma_j^2 + \sum_{i:y^i=1}\sum_{j=1}^k -\frac{1}{2\sigma_j^2}(x_j^i - \mu_{1,j})^2 + \sum_{i:y^i=0}\sum_{j=1}^k -\frac{1}{2\sigma_j^2}(x_j^i - \mu_{0,j})^2 \\ &\quad + N_1\log\phi + N_0\log(1-\phi) + \text{const.} \\ &= -\frac{N}{2}\sum_{j=1}^k \log\sigma_j^2 + \sum_{j=1}^k -\frac{1}{2\sigma_j^2}\left(\sum_{i:y^i=1}(x_j^i - \mu_{1,j})^2 + \sum_{i:y^i=0}(x_j^i - \mu_{0,j})^2\right) \\ &\quad + N_1\log\phi + N_0\log(1-\phi) + \text{const.}\end{aligned}$$

In the above, x_j^i denotes the j-th element of $\boldsymbol{x}^i$ and similarly for $\mu_{1,j}$ and $\mu_{0,j}$. Taking the derivative of the above with respect to σ_j^2 and setting it to zero, we obtain

$$\hat{\sigma}_j^2 = \frac{1}{N}\left(\sum_{i:y^i=1}(x_j^i - \hat{\mu}_{1,j})^2 + \sum_{i:y^i=0}(x_j^i - \hat{\mu}_{0,j})^2\right)$$

10) Linear Regression Implementation:
a) Write down a code in Python whose input is a training dataset $\{(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^N, y^N)\}$ and its output is the weight vector $\boldsymbol{\theta}$ in the linear regression model $y = \boldsymbol{\theta}^\top \phi(\boldsymbol{x})$, for a given nonlinear

mapping $\phi(\cdot)$. Implement two cases: i) using the closed-form solution, ii) using a stochastic gradient descent on mini-batches of size m.

b) Consider n-degree polynomials, $\phi(\cdot) = \begin{bmatrix} 1 & x & x^2 & \cdots & x^n \end{bmatrix}$. Download the dataset on the course webpage and work with 'dataset1'. Run the code on the training data to compute $\boldsymbol{\theta}$ for $n \in \{2, 3, 5\}$. Evaluate the regression error on both training and the test data. Report $\boldsymbol{\theta}$, training error and test error for both implementation (closed-form vs gradient descent). What is the effect of the size of the mini-batch on the speed and testing error of the solution.

c) Download the dataset on the course webpage and work with 'dataset2'. Write a code in Python that applies Ridge regression to the dataset to compute $\boldsymbol{\theta}$ for a given λ. Implement two cases: using a closed-form solution and using a stochastic gradient descent method with mini-batches of size m. Use K-fold cross validation on the training dataset to obtain the best regularization λ and apply the optimal $\boldsymbol{\theta}$ to compute the regression error on test samples. Report the optimal λ, $\boldsymbol{\theta}$, test and training set errors for $K \in \{2, 10, N\}$, where N is the number of samples. In all cases try $n \in \{2, 3, 5\}$. How does the test error change as a function of λ and n?