New York University Tandon School of Engineering
Computer Science and Engineering

CS-GY 6923: Written Homework 2.
Due Tuesday, October 15th, 2024, 11:59pm.

NO SLIP DAY FOR THIS HOMEWORK.

Discussion with other students is allowed for this problem set, but solutions must be written-up individually.

Problem 1: Impacts of Regularization (10pts)

Consider the ridge regularized least squares regression problem $\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$ with different positive values of λ. Let $\boldsymbol{\beta}_1^* = \arg\min \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_2^2$ and $\boldsymbol{\beta}_2^* = \arg\min \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2$.

(a) Prove that if $\lambda_1 \geq \lambda_2$ then $\|\boldsymbol{\beta}_1^*\|_2^2 \leq \|\boldsymbol{\beta}_2^*\|_2^2$. In words, increasing the regularization parameter *always* decreases the norm of the optimal parameter vector.

(b) Prove that if $\lambda_1 \geq \lambda_2$ then $\|\mathbf{X}\boldsymbol{\beta}_1^* - \mathbf{y}\|_2^2 \geq \|\mathbf{X}\boldsymbol{\beta}_2^* - \mathbf{y}\|_2^2$. In words, increasing the regularization parameter *always* leads to higher training loss, even if it might improve test loss.

(c) Suppose instead that we used LASSO regularization, so that Let $\boldsymbol{\beta}_1^* = \arg\min \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1$ and $\boldsymbol{\beta}_2^* = \arg\min \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_1$. Do the above conclusions change?

Problem 2: Gaussian Naive Bayes (20pts)

The Naive Bayes Classifier can be extended to predictor variables with continuous values (instead of just binary variables). Consider a data set where each example $(\mathbf{x}, y)$ contains a data vector $\mathbf{x} \in \mathbb{R}^d$ and a label $y \in \{0, 1\}$. As in class, each y is modeled as a Bernoulli random variable, which equals 1 with probability p and 0 with probability $1 - p$. To model $\mathbf{x}$ we have two lists of mean/variances pairs:

$$(\mu_{0,1}, \sigma_{0,1}^2), (\mu_{0,2}, \sigma_{0,2}^2), \ldots, (\mu_{0,d}, \sigma_{0,d}^2) \qquad \text{and} \qquad (\mu_{1,1}, \sigma_{1,1}^2), (\mu_{1,2}, \sigma_{1,2}^2), \ldots, (\mu_{1,d}, \sigma_{1,d}^2).$$

If y equals 0, then the j^{th} entry of $\mathbf{x}$ is modeled as an *independent* Gaussian random variable with mean $\mu_{0,j}$ and variance $\sigma_{0,j}^2$. Alternatively, if y equals 1, then the j^{th} entry of $\mathbf{x}$ is modeled as an independent Gaussian random variable with mean $\mu_{1,j}$ and variance $\sigma_{1,j}^2$.

(a) Given a training data set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ write down expressions for estimating all model parameters $\mu_{i,j}$ and $\sigma_{i,j}^2$ from the data. (This is a relatively simple question).

(b) Given a new unlabled predictor vector $\mathbf{x}_{new}$ we would like to predict class label y_{new} using a *maximum a posterior* (MAP) estimate. In other words, we want to choose y_{new} to maximize the posterior probability $p(y_{new} \mid \mathbf{x}_{new})$. Write pseudocode for determining if $p(y_{new} = 0 \mid \mathbf{x}_{new})$ or $p(y_{new} = 1 \mid \mathbf{x}_{new})$ is larger. **Hint:** You probably want to use Bayes rule!

(c) Implement your method by completing the Python workbook hw2_stub.ipynb. Attach a printed PDF of your completed notebook results to your homework submission. To avoid underflow issues, you might want to work with the log of the probabilities instead of the probabilities directly – i.e., your code should target the problem of determining $\log\left(p(y_{new} = 0 \mid \mathbf{x}_{new})\right)$ or $\log\left(p(y_{new} = 1 \mid \mathbf{x}_{new})\right)$ is larger.

(d) Consider the probabilistic model above with $d = 1$. So, our dataset consists of (x, y) pairs where x is a scalar value. Suppose $p(x \mid y = 1)$ is a Gaussian pdf with $\mu = 2$ and $\sigma = 3$, and that $p(x \mid y = 0)$ is also Gaussian with $\mu = 5$ and $\sigma = 3$. Suppose, too, that $P(y = 1) = P(y = 0) = \frac{1}{2}$.

Suppose we implement the MAP prediction rule you developed above for this dataset. Calculate the probability that the prediction will be incorrect.

Hint: Your solution will likely require computing the area under a Gaussian pdf. There are a number of ways to do this. For example, you can use something like `scipy.stats.norm.cdf`. We will let you read the documentation to see how to use it, but it may be helpful to remember that if F is the cumulative density function for a distribution with density f, then $\int_a^b f(x)\,dx = F(b) - F(a)$.

Problem 3: More Practice with MAP Calculations (5pts)

Consider a model similar to the one above, but with a non-Gaussian distribution. In particular, each data points has a binary class label $y \in \{0, 1\}$. We have prior class probabilities: $P(y = 0) = .4$ and $P(y = 1) = .6$. The data x, conditioned on the class labels y, is modeled as a continuous random variable that takes values on the interval $[0, 7]$. Specifically, the conditional densities for x are known to be:

$$p(x|y=0) = \begin{cases} \frac{1}{5}, & \text{for } 0 \le x \le 2 \\ \frac{1}{3}, & \text{for } 2 < x \le 3 \\ \frac{1}{15}, & \text{for } 3 < x \le 7 \end{cases}$$

$$p(x|y=1) = \begin{cases} \frac{1}{6}, & \text{for } 0 \le x \le 1 \\ \frac{1}{8}, & \text{for } 1 < x \le 5 \\ \frac{1}{6}, & \text{for } 5 < x \le 7 \end{cases}$$

Given a new data point x_{new}, suppose we use Bayes' theorem to compute the posterior probability of each class given x_{new}, and make a predication based on the maximum posterior (i.e., use a MAP estimate). For what values of $x \in [0, 7]$ will this approach predict $y = 1$?

Problem 4: Bayesian Central Tendency (9pts)

Let's revisit a question on the first homework from a Bayesian perspective.

(a) Suppose we have a data set of scalar numbers $x_1, \ldots, x_n$. Assume a Bayesian probabilistic model in which the numbers are drawn from a Gaussian distribution with unknown mean μ and variance σ^2. We have no prior information on μ and σ^2: we assume all parameters are equally likely. Prove that the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an MLE estimate for the unknown parameter μ. I.e $\hat{\mu} = \arg\max_{\mu} \Pr(x_1, \ldots, x_n \mid \mu)$.

(b) Now assume a Bayesian probabilistic model in which the numbers are drawn from a Laplace Distribution with unknown mean μ and variance $2b^2$. Prove that the sample median is a MLE estimate for the unknown parameter μ.

(c) Suppose $\mu \in [0, 1]$ and $x_1, \ldots, x_n$ are drawn i.i.d from a Bernoulli distribution with parameter μ. I.e. x_i is 1 with probability μ and 0 with probability $1 - \mu$. Prove that the sample mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is also an MLE estimator for μ in this setting.

Problem 5: Unexpected Mean Estimators (10pts)

In the problem above, the Bayesian approach suggested very intuitive and familiar estimates for the mean – the sample average, sample median, etc. However, for some distributions, much more unusual mean estimators turn out to be more effective.

The Rayleigh distribution is often used to model the magnitude of vectors in 2D space where the components are independent Gaussian random variables. Specifically, for a random 2D vector $\mathbf{v} = (x, y)$, where $x \sim \mathcal{N}(0, \sigma^2)$ and $y \sim \mathcal{N}(0, \sigma^2)$, the magnitude $r = \sqrt{x^2 + y^2}$ follows a Rayleigh distribution. This distribution frequently arises in applications such as wireless communication (signal strength in multipath environments) and medical imaging (ultrasound speckle noise).

The Rayleigh distribution has probability density function (pdf):

$$p(x) = \frac{x}{\sigma^2} \mathrm{e}^{-x^2/(2\sigma^2)}, \tag{1}$$

where $\sigma \cdot \sqrt{\pi/2}$ is the mean. Given this, a natural approach to estimating the unknown parameter σ from data drawn from a Rayleigh distribution could be to compute the sample mean and multiply by $\sqrt{\pi/2}$. As you will show, this is surprisingly suboptimal. The maximum likelihood estimator looks quite different.

a) Consider data points $x_1, \ldots, x_n$ drawn independent from a Rayleigh distribution with parameter σ. Write down an expression for the log-likelihood of $x_1, \ldots, x_n$ given the parameter σ. I.e., write down an expression for $\ln(p(x_1, \ldots, x_n \mid \sigma))$ where ln denotes the natural logarithm (base e).

b) Using your expression above, show that the maximum likelihood estimate of σ is

$$\sigma_{MLE} = \sqrt{\frac{1}{2n}\sum_{i=1}^{n} x_i^2}. \tag{2}$$