Optimization and Machine Learning, Spring 2020
Homework 2

(Due Wednesday, Apr. 1 at 11:59pm (CST))

March 27, 2020

1. Suppose that we have N training samples, in which each sample is composed of p input variable and one categorical response with K states.
   (a) Please define this multi-class classification problem, and solve it by ridge regression. (4 points)
   (b) Please make the prediction of a testing sample $x \in \mathbb{R}^p$ based on your model in (a). (3 points)
   (c) Is there any limitation on your model? If yes, please explain the problem by drawing a picture. (3 points)
   (d) Can you propose a model to overcome this limitation? If yes, please derive the decision boundary between an arbitrary class-pair. (5 points)
   (e) Can you revise your model in (d) by strength or weaken its assumptions? If yes, please tell the difference between your models in (d) and (e). (5 points)

2. Given an random variable, we have N i.i.d. observations by repeated experiments.
   (a) If the variable is boolean, please calculate the log-likelihood function. (4 points)
   (b) If the variable is categorical, please calculate the log-likelihood function. (4 points)
   (c) If the variable is continuous and follows Gaussian distribution, please calculate the log-likelihood function. (5 points)
   (d) Please discuss the difference between Maximum Likelihood Estimation (MLE) and Maximum a Posterior (MAP) estimation based on ONE of your results in (a), (b) and (c). (7 points)

3. Given the input variables $X \in \mathbb{R}^p$ and a response variable $Y \in \{0, 1\}$, the Expected Prediction Error (EPE) is defined by

$$\text{EPE} = \mathbb{E}[L(Y, \hat{Y}(X))],$$

   where $\mathbb{E}(\cdot)$ denotes the expectation over the joint distribution $\Pr(X, Y)$, and $L(Y, \hat{Y}(X))$ is a loss function measuring the difference between the estimated $\hat{Y}(X)$ and observed Y.
   (a) Given the zero-one loss

$$L(k, \ell) = \begin{cases} 1 & \text{if } k \neq \ell \\ 0 & \text{if } k = \ell, \end{cases}$$

   please derive the Bayes classifier $\hat{Y}(x) = \text{argmax}_{k \in \{0,1\}} \Pr(Y = k | X = x)$ by minimizing EPE. (2 points)
   (b) Please define a function which enables to map the range of an arbitrary linear function to the range of a probability. (2 points)
   (c) Based on the function you defined in (b), please approximate the Bayes classifier in (a) by a linear function between X and Y, and derive its decision boundary. (4 points)
   (d) If each element of X is boolean, please show how many independent parameters are needed in order to estimate $\Pr(Y|X)$ directly; and is there any way to reduce its number? If yes, please describe your way mathematically. (4 points)
   (e) Based on your results in (d) and the Bayes theorem, please develop a classifier with a linear number of parameters w.r.t. p, and estimate these parameters by MLE. (5 points)
   (f) Please find at least three different points between your developed models in (c) and (e). (3 points)

4. Consider 12 labeled data points sampled from three distinct classes:

$$\text{Class } 0: \begin{bmatrix}0\\2\end{bmatrix}, \begin{bmatrix}-2\\0\end{bmatrix}, \begin{bmatrix}5\\3\end{bmatrix}, \begin{bmatrix}-3\\-5\end{bmatrix} \quad \text{Class } 1: \begin{bmatrix}\sqrt{2}\\\sqrt{2}\end{bmatrix}, \begin{bmatrix}-\sqrt{2}\\\sqrt{2}\end{bmatrix}, \begin{bmatrix}4\sqrt{2}\\-\sqrt{2}\end{bmatrix}, \begin{bmatrix}-4\sqrt{2}\\-\sqrt{2}\end{bmatrix}, \quad \text{Class } 2: \begin{bmatrix}3\\5\end{bmatrix}, \begin{bmatrix}-1\\3\end{bmatrix}, \begin{bmatrix}8\\6\end{bmatrix}, \begin{bmatrix}0\\-2\end{bmatrix}$$

   (a) For each class $C \in [0, 1, 2]$, compute the class sample mean μ_C, the class sample covariance matrix Σ_C, and the estimate of the prior probability π_C that a point belongs to class C. (6 points)

   (b) Suppose that we apply LDA to classify the data given in part (a). Will this get the good decision boundary? Briefly explain your answer. (4 points)

5. We have two classes, named N for normal and E for exponential. For the former class ($Y = N$), the prior probability is $\pi_N = P(Y = N) = \frac{\sqrt{2\pi}}{1+\sqrt{2\pi}}$ and the class conditional $P(X|Y = N)$ has the normal distribution $N(0, \sigma^2)$. For the latter, the prior probability is $\pi_E = P(Y = E) = \frac{1}{1+\sqrt{2\pi}}$ and the class conditional has the exponential distribution.

$$P(X = x|Y = E) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

   Write an equation in x for the decision boundary. (Only the positive solutions of your equation will be relevant; ignore all $x < 0$.) Simplify the equation until it is quadratic in x. (You don't need to solve the quadratic equation. It should contain the constants σ and λ. Ignore the fact that 0 might or might not also be a point in the decision boundary.) (10 points)

6. Given data $\{(x_i, y_i) \in R^d \times \{0, 1\}\}_{i=1}^n$ and a query point x, we choose a parameter vector θ to minimize the loss (which is simply the negative log likelihood, weighted appropriately):

$$l(\theta; x) = -\sum_{i=1}^{n} w_i(x)[y_i \log(\mu(x_i)) + (1 - y_i) \log(1 - \mu(x_i))]$$

   where

$$\mu(x_i) = \frac{1}{1 + e^{-\theta \cdot x_i}}, w_i(x) = \exp(-\frac{||x - x_i||^2}{2\tau})$$

   where τ s a hyperparameter that must be tuned. Note that whenever we receive a new query point x, we must solve the entire problem again with these new weights $w_i(x)$.

   (a) Given a data point x, derive the gradient of $l(\theta; x)$ with respect to θ. (4 points)

   (b) Given a data point x, derive the Hessian of $l(\theta; x)$ with respect to θ. (4 points)

   (c) Given a data point x, write the update formula for Newton's method. (2 points)

7. Now we discuss Bayesian inference in coin flipping. Let's denote the number of heads and the total number of trials by N_1 and N, respectively.

   (a) Please derive the MAP estimation based on the prior $p(\theta) = \text{Beta}(\theta|\alpha, \beta)$. (4 points)

   (b) Please derive the MAP estimation based on the following prior:

$$p(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.5 \\ 0.5 & \text{if } \theta = 0.4 \\ 0 & \text{otherwise,} \end{cases}$$

   that believes the coin is fair, or is slightly biased towards tails. (4 points)

   (c) Suppose the true parameter is $\theta = 0.41$. Which prior leads to a better estimate when N is small? Which prior leads to a better estimate when N is large? (2 points)