CS5487: Take-home quiz

2019 Semester A

Dec 12 to Dec 18

Rules:

1. This take-home quiz is an "open-book" quiz. You are permitted to use the following materials during the quiz:

   - Your lecture notes.
   - Your cheatsheet from the Midterm Quiz.
   - The textbook, Pattern Recognition and Machine Learning (PRML) by Bishop.
   - Any materials available on the CS5487 Canvas course page, including Problem Sets, Problem Set Solutions, Tutorial Solutions, Panopto Recordings.

   All other materials are NOT allowed. This includes web searches, research papers, and other reference books, etc.

2. You cannot discuss the quiz with others, and the work that you turn in must be your own work. You will follow the high standards of Academic Honesty at CityU.

3. You have until Dec 18, 5pm to complete the quiz. Turn in your work on Canvas.

Instructions:

1. Answer all questions on blank paper.

2. On the last page of your answer sheets, write the following statement: "The work in these answer sheets are my own work. I have not discussed this quiz with anyone else. I have only used the allowed materials." Then write your name, student number, date and put your signature.

3. Upload your answer sheets to Canvas.

Problem 1 Soft Adaptive-SVM [30 marks]

In this problem we will consider an adaptive-SVM (ASVM) for binary classification. Suppose we have used a dataset $\mathcal{D}_0$ to learn a binary linear classifier function $f_0(x) = w_0^T x$ with decision rule $y = \text{sign}(f_0(x))$. Since we have the classifier, we then discarded the data $\mathcal{D}_0$.

Now, suppose we receive a new set of data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ are the feature vectors and $y_i \in \{+1, -1\}$ the corresponding class. We wish to update our original classifier function $f_0(x)$. To do this, we will add a "delta classifier" $\Delta f(x) = w^T x$ to adapt our original classifier $f_0(x)$ into a new classifier $f(x)$,

$$f(x) = f_0(x) + \Delta f(x) = f_0(x) + w^T x, \tag{1}$$

where w is the parameter vector of the "delta classifier". To handle cases when the data is not linearly separable, we introduce slack variables ξ_i for each data point x_i. The ASVM primal problem is

$$\begin{aligned} &\min_w \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \\ &\text{s.t. } y_i(f_0(x_i) + w^T x_i) \geq 1 - \xi_i, \quad \forall i, \\ &\quad\ \xi_i \geq 0, \quad \forall i. \end{aligned} \tag{2}$$

(a) [2 marks] Explain the role of the objective function and the constraints in the ASVM primal problem.

(b) [5 marks] Write down the Lagrangian $L(w, \xi, \alpha, r)$ for the ASVM primal problem, where α are the Lagrange multipliers for the first set of inequality constraints, and r are the Lagrange multipliers for the second set of inequality constraints. Derive conditions for the stationary point of $L(w, \xi, \alpha, r)$ w.r.t. w and ξ.

(c) [10 marks] Derive the ASVM dual problem.

(d) [3 marks] Use the KKT conditions to derive a geometric interpretation of the ASVM.

(e) [10 marks] Compare the ASVM dual in (c) with the original soft-SVM dual problem. What is the interpretation of the ASVM dual (considering the original SVM dual)? What is the role of the original classifier $f_0(x)$?

....

Problem 2 Gaussian variance regression [50 marks]

Consider the regression problem where $x \in \mathbb{R}^d$ is the input vector and $y \in \mathbb{R}$ is the observation value. The training set is $\mathcal{D} = \{X, y\}_{i=1}^n$, where $X = [x_1, \cdots, x_n]$ are the input vectors, and $y = [y_1, \cdots, y_n]^T$ are the output values.

In this problem, we will consider a Gaussian observation model with fixed mean $\mu = 0$ and variance σ^2 that changes as a function of x. That is, our goal is to regress the variance of the Gaussian using the inputs X and the corresponding observations y. The Gaussian observation likelihood with mean 0 is

$$p(y|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}y^2}. \tag{3}$$

Since the variance should be non-negative, we define the mapping from x to the variance σ^2 as the exponential of a linear function

$$\sigma^2(x) = e^{-w^T x}, \tag{4}$$

where $w \in \mathbb{R}^d$ is the parameter vector. Thus, the observation likelihood in terms of w, x is given by

$$p(y|w, x) = \frac{1}{\sqrt{2\pi(e^{-w^T x})}} e^{-\frac{1}{2}(e^{w^T x})y^2}, \tag{5}$$

We also assume a Gaussian prior on the weight vector w,

$$p(w) = \mathcal{N}(w|0, \Sigma). \tag{6}$$

First we will consider the MAP estimate of the regression parameters w.

(a) [5 marks] Describe a real-world problem where this type of regression could be used.

(b) [5 marks] Write down the optimization problem for the MAP estimate of w.

(c) [10 marks] Derive the Newton-Raphson iterations to solve for the MAP estimate of w.

(d) [5 marks] Consider the case when the prior covariance matrix is $\Sigma = \lambda I$. How does Σ help to regularize the estimate of w?

Now we will consider a non-linear version by kernelizing the regression model.

(e) [5 marks] Derive the *kernel* version of the regression model, i.e., let $\sigma_*^2 = e^{-\alpha_*}$, and apply the *kernel trick* to calculate $\alpha_* = w^T x_*$.

(f) [10 marks] Derive the kernel version of the MAP estimation using the Newton-Raphson iterations derived in (c).

(g) [5 marks] Discuss the role of the prior covariance Σ in the kernel regression model.

(h) [5 marks] Compare the original and kernelized algorithms in (c) and (f). What are the advantages and disadvantages of each version?

....