COMPSCI 671D Fall 2019
Homework 3

Total of 85 points.

1 Kernels (25 points)

Suppose we have two valid kernels k_1, k_2, prove/disprove that the following kernels are valid. Do not just quote from the lecture notes that they are valid, you'll actually need to provide a proof. Please try it yourself before looking it up on the Internet.

(a) (3 points)

$$k(x, z) = \alpha k_1(x, z) + \beta k_2(x, z), \text{ for } \alpha, \beta \geq 0$$

(b) (3 points)

$$k(x, z) = k_1(x, z)k_2(x, z)$$

(c) (3 points)

$$k(x, z) = f(x)f(z) \text{ for } f : \mathcal{X} \to \mathbb{R}$$

(d) (3 points)

$$k(x, z) = f(k_1(x, z)), \text{ where } f(x) \text{ is a polynomial with positive coefficients.}$$

(e) (3 points)

$$k(x, z) = \exp(-\gamma \|x - z\|_2^2)$$

(f) (5 points) Show that the **rbf** kernel for some fixed $\gamma \in \mathbb{R}^+$ corresponds to the inner product of some feature map ϕ in infinite dimensional space, that is,

$$k(x, z) = \exp(-\gamma \|x - z\|_2^2) = \phi(x)^T \phi(z)$$

where $x, z \in \mathcal{X}, \quad \phi : \mathcal{X} \to \mathcal{X}^\infty$
(Hint: Taylor Expansion)

(g) (5 points) Suppose we have a dataset $\{x_i, y_i\}_{i=1}^n, x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}$ with no conflicting labels, where a conflicting label means $\exists i, j \quad s.t. \quad x_i = x_j, y_i \neq y_j$. Show that a SVM classifer with the **rbf** kernel can always achieve 0 training error.
(Hint: Consider what happens when $\gamma \to \infty$)

2 Statistical Learning Theory (25 points)

(a) (7 points) Prove that the VC-dimension of affine classifiers (half spaces) in $\mathbb{R}^d$ is $d+1$.

(b) (5 points) Find the VC-dimension of SVM with polynomial kernel $k(x,z) = (x^T z+1)^d, x, z \in \mathbb{R}^p$ and prove your result.

(c) (3 points) Find the VC-dimension of SVM with **rbf** kernel $k(x,z) = \exp(-\gamma\|x-z\|_2^2)$, for some fixed $\gamma \in \mathbb{R}^+$, where $x, z \in \mathbb{R}^p$ and prove your result.

(d) (10 points) Suppose we have a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$ with $\mathbf{x}_i \in \mathcal{X} = \{0,1\}^p, y_i \in \mathcal{Y} = \{0,1\}$. Consider constructing a function $f : \mathcal{X} \to \mathcal{Y}$ to predict y from $\mathbf{x}$.

(i) If we set the function class $\mathcal{F}$ to be the conjunction of at most p literals, derive an upper bound of the true risk $R^{\text{true}}(f)$ for any $f \in \mathcal{F}$ at confidence level $1-\delta$. Represent your upper bound with the empirical risk $R^{\text{emp}}(f)$, δ and other quantities given in the problem. (Note: A literal of a boolean variable x is either x or $\neg x$. A conjunction of k variables is $x_1 \wedge x_2 \cdots \wedge x_k$. Assume we do not take the conjunction of positive and negative literals of the same variable, that is $x \wedge \neg x$.)

(ii) What would the upper bound be if we set the function class $\mathcal{F}$ to be the conjunction of at most p' literals, where $p' \leq p$.

3 Ridge Regression (20 points)

Given a dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, suppose the data is generated as follows:

$$y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i, \text{ where } \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\lambda}\mathbf{I}_p) \text{ and noise } \epsilon_i \sim \mathcal{N}(0,1),$$

where y_i are conditionally independent given $\mathbf{w}$. We can stack our data in a $n \times p$ matrix $\mathbf{X}$, and our labels in a $n \times 1$ vector $\mathbf{y}$.

(a) (4 points) Please give an expression for the likelihood $\Pr(\mathbf{y}|\mathbf{X}, \mathbf{w})$.

(b) (8 points) Please give the Maximum A Posteriori (MAP) expression for $\mathbf{w}$.

Hint: first show the following:

$$\Pr(\mathbf{w}|\mathbf{y}, \mathbf{X}, \lambda) \propto \exp\left\{-\frac{1}{2}\left[\mathbf{w}^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} - 2\mathbf{w}^T\mathbf{X}\mathbf{y}\right]\right\}$$

$$\Pr(\mathbf{w}|\mathbf{y}, \mathbf{X}, \lambda) \propto \exp\left\{-\frac{1}{2}(\mathbf{w} - (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y})^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})(\mathbf{w} - (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y})\right\}.$$

(c) (4 points) We can also obtain an estimate of $\mathbf{w}$ by minimizing the L^2 loss with L^2 regularisation.

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2, \tag{1}$$

where λ is a hyperparameter that controls the amount of regularization.
Give the expression for $\mathbf{w}$ that minimizes the above objective function.

(d) (1 point) From your answer in (b) and (c) what are the dual interpretations of ridge regression?

(e) (3 points) Suppose our generative model is wrong: $\mathbf{w}$ is not distributed with mean $\mathbf{0}$. Does adding $\lambda\|\mathbf{w}\|^2$ as the regularization term introduce additional bias? If so, (i) How may you preprocess and/or modify the objective function to reduce this additional bias? Assume you know the correct mean for the generative process for $\mathbf{w}$. (ii) If we do not know the correct mean, but that we know $\mathbf{w}$ is not distributed with mean $\mathbf{0}$, why should we add any regularization?

4 Catch Mice with Clustering (15 points)

Suppose you were given m mouse tracks, each track is of duration p, i.e. $\{\mathbf{m}_i\}_{i=1}^m, \mathbf{m}_i \in \mathbb{R}^{p\times 2}$. We are asked to place K mouse traps to catch some mice. A possible solution would be to optimize the placement of mouse traps to minimize the distances between trap locations and all of the locations where a mouse has been spotted so far:

$$\min_{\mathbf{c}_1,\ldots,\mathbf{c}_K} \sum_{i=1}^{N} \min_{k\in\{1,\ldots,K\}} \|\mathbf{x}_i - \mathbf{c}_k\|_1,$$

where $\{\mathbf{c}_k\}_{k=1}^K, \mathbf{c}_k \in \mathbb{R}^2$ are cluster centers representing the locations of mouse traps, and $\{\mathbf{x}_i\}_{i=1}^{N=m\times p}$, $\mathbf{x}_i \in \mathbb{R}^2$ are data points representing the locations of mice, and $\|\cdot\|_1$ is the L^1 norm, also known as the taxicab norm or Manhattan norm.

(a) (1 point) How is the above objective function different from k-means' objective taught in class?

(b) (6 points) Implement k-means taught in class on the mice location data, sample .pynb code to load the data is provided.

(c) (8 points) Using a similar derivation to k-means taught in class, derive the cluster assignment and cluster update steps that minimize the above objective function. Then implement this algorithm on the mouse location data. This algorithm is also known as k-medians.

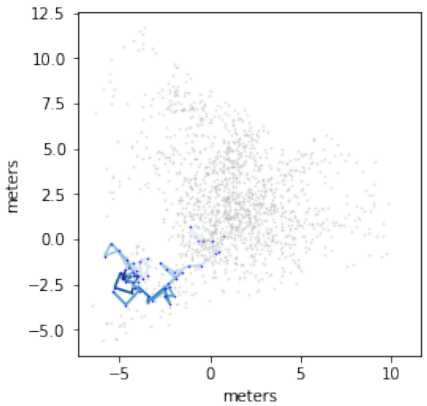

Figure 1: Mouse tracks from 30 mice. A single mouse track has been highlighted in blue.