CS 189 Introduction to Machine Learning
Summer 2019 Marc Khoury & Brijen Thananjeyan

DIS10

1 Kernelizing Nearest Neighbors

In this question, we will be looking at how we can kernelize k-nearest neighbors (k-NN). k-NN is a simple classifier that relies on nearby sample points to decide what a new point's class should be.

Given a query point q, there are 2 steps to decide what class to predict.

1. Find the k sample points nearest q.
2. Return the class with the most votes from the k sample points.

For the following parts, assuming that our sample points $x \in \mathbb{R}^d$, and we have n sample points.

(a) What is the runtime to classify a newly given query point q, using euclidean distance?

(b) What if instead of looking at the distance between the points in $\mathbb{R}^d$, we wanted to consider the distance in p polynomial space? What dimension would this space be? What would the runtime be to classify a newly given query point q, in terms of n, d, and k, and p?

(c) Instead, we can use the polynomial kernel to compute the distance between 2 points in p polynomial space without having to move all of the points into the higher dimensional space. Using the polynomial kernel, $k(x, y) = (x^T y + \alpha)^p$ instead of Euclidean distance, what is the runtime for k-NN to classify a new point q?

2 Gaussian Kernels

In this question, we will look at training a binary classifier with a Gaussian kernel. Specifically given a labelled dataset $S = \{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \{\pm 1\}$ and a kernel function $k(x_1, x_2)$, we consider classifiers of the form:

$$\widehat{f}(x) = \text{sign}\left(\sum_{i=1}^{n} a_i k(x_i, x)\right),$$

where we define sign(u) to be 1 if $u \geq 0$ or -1 if $u < 0$. In order to choose the weights $a_i, i = 1, ..., n$, we will consider the least-squares problem:

$$a \in \arg\min_{a \in \mathbb{R}^n} \|Ka - y\|_2^2, \tag{1}$$

where $K = (k(x_i, x_j))_{i=1,j=1}^{n}$ is the kernel matrix and $y = (y_1, ..., y_n)$ is the vector of labels. We will work with the Gaussian kernel. Recall that the Gaussian kernel with bandwidth $\sigma > 0$ is defined as:

$$k(x_i, x_j) := \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right).$$

(a) When the bandwidth parameter $\sigma \to 0$, observe that the off-diagonal entries of the kernel matrix K tend also to zero. Consider a two sample dataset S (i.e. $n = 2$) with $(x_1, y_1) = (1, 1)$ and $(x_2, y_2) = (-1, -1)$. Assuming that as $\sigma \to 0$ the off-diagonal entries of K are equal to zero (and the diagonal entries unmodified), what is the optimal solution of a for the optimization problem (1) and what is the resulting classifier $\widehat{f}(x)$?

(b) Now we consider the regime when the bandwidth parameter $\sigma \to +\infty$. Observe in this regime, the off-diagonal entries of the kernel matrix K tend to one. Given a dataset S, suppose we solve the optimization problem (1) with all the off-diagonal entries of K equal to one (and the diagonal entries unmodified). Prove that if the number of $+1$ labels in S equals the number of -1 labels in S, then $a = \mathbf{0}$ is an optimal solution of (1). What is the resulting classifier $\widehat{f}(x)$?

(c) Now we consider the regime when the bandwidth parameter is large but finite. Consider again the two sample dataset S with $(x_1, y_1) = (1, 1)$ and $(x_2, y_2) = (-1, -1)$. When $\sigma \gg 1$, we can approximate $k(x_1, x_2) \approx 1 + \frac{x_1 x_2}{2\sigma^2}$. Show that the solution of the optimization problem (1) with the kernel $k_a(x_1, x_2) = 1 + \frac{x_1 x_2}{2\sigma^2}$ is given by $a = (\sigma^2, -\sigma^2)$. What is the resulting classifier $\widehat{f}(x)$?

Hint: The inverse of a 2×2 *matrix is given by the formula* $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$.

3 Kernel Validity

For a function $k(x_i, x_j)$ to be a valid kernel, it suffices to show either of the following conditions is true:

1. k has an inner product representation: $\exists\ \Phi : \mathbb{R}^d \to \mathcal{H}$, where $\mathcal{H}$ is some (possibly infinite-dimensional) inner product space such that $\forall x_i, x_j \in \mathbb{R}^d,\ k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$.
2. For every sample $x_1, x_2, \ldots, x_n \in R^d$, the kernel matrix

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & k(x_i, x_j) & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix}$$

is positive semidefinite. For the following parts you can use either condition (1) or (2) in your proofs.

(a) Show that the first condition implies the second one, i.e. if $\forall x_i, x_j \in \mathbb{R}^d,\ k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$ then the kernel matrix K is PSD.

(b) Given a positive semidefinite matrix A, show that $k(x_i, x_j) = x_i^\top A x_j$ is a valid kernel.

(c) Show why $k(x_i, x_j) = x_i^\top(\text{rev}(x_j))$ (where rev(x) reverses the order of the components in x) is *not* a valid kernel.

(d) When solving Kernel ridge regression, one can show that the key intermediate step is solving the following optimization problem:

$$\text{argmin}_{\alpha \in \mathbb{R}^n} \left[\frac{1}{2} \alpha^T (K + \lambda I)\alpha - \lambda \langle \alpha, y \rangle \right]$$

where $y \in \mathbb{R}^n$, $\lambda \geq 0$, and $K \in \mathbb{R}^{n \times n}$ is the kernel matrix computed by applying a kernel function k on every sample pair: $k(x_i, x_j)$. When λ is close to 0, why is it important that K is a valid kernel?

4 Polynomial Kernel

An alternative formulation of the polynomial kernel is

$$k(x, y) = (x^T y + \alpha)^p$$

where $x, y \in \mathbb{R}^n$, and $\alpha \geq 0$. When we take $p = 2$, this kernel is called the quadratic kernel. Find the feature mapping $\Phi(z)$ that corresponds to the quadratic kernel.