Foundations of Machine Learning

Brett Bernstein

August 22, 2018

Week 5 Lab: Concept Check Exercises

Kernels

Kernel Learning Objectives

- Explain how explicit feature maps can be used to extend the expressivity of linear models.
- Explain potential issues explicitly computing large feature spaces.
- State and explain the definition of a 'kernelized' method.
- Explain why the SVM dual is kernelized, while the primal is not (ignoring the representer theorem).
- Give the relationship between a feature map and kernel function.
- Explain the computational benefits of kernelization based on costs of optimizing over $\mathbb{R}^n$ vs $\mathbb{R}^d$.
- Be able to apply the kernel trick using the kernel matrix K.
- Be able to apply the elements of our proof of the representer theorem (ex: projections decrease norms) to prove related theorems.
- Compare using the representer theorem and duality to kernelized SVM.
- Describe common kernels (RBF/polynomial) and their properties (i.e. equivalent feature maps, computational benefits relative to explicit computation (if possible),...).
- Describe some general recipes for deriving "new" kernel function.

Kernel Concept Check Questions

1. Fix $n > 0$. For $x, y \in \{1, 2, \ldots, n\}$ define $k(x, y) = \min(x, y)$. Give an explicit feature map $\varphi : \{1, 2, \ldots, n\}$ to $\mathbb{R}^D$ (for some D) such that $k(x, y) = \varphi(x)^T \varphi(y)$.

2. Show that $k(x, y) = (x^T y)^4$ is a positive semidefinite kernel on $\mathbb{R}^d \times \mathbb{R}^d$.

3. Let $A \in \mathbb{R}^{d\times d}$ be a positive semidefinite matrix. Prove that $k(x, y) = x^T Ay$ is a positive semidefinite kernel.

4. Consider the objective function

$$J(w) = \|Xw - y\|_1 + \lambda \|w\|_2^2.$$

   Assume we have a positive semidefinite kernel k.

   (a) What is the kernelized version of this objective?

   (b) Given a new test point x, find the predicted value.

5. Show that the standard 2-norm on $\mathbb{R}^n$ satisfies the parallelogram law.

6. Suppose you are given an training set of distinct points $x_1, x_2, \ldots, x_n \in \mathbb{R}^n$ and labels $y_1, \ldots, y_n \in \{-1, +1\}$. Show that by properly selecting σ you can achieve perfect $0-1$ loss on the training data using a linear decision function and the RBF kernel.

7. Suppose you are performing standard ridge regression, which you have kernelized using the RBF kernel. Prove that any decision function $f_\alpha(x)$ learned on a training set must satisfy $f_\alpha(x) \to 0$ as $\|x\|_2 \to \infty$.

8. Consider the standard (unregularized) linear regression problem where we minimize $L(w) = \|Xw - y\|_2^2$ for some $X \in \mathbb{R}^{n\times m}$ and $y \in \mathbb{R}^n$. Assume $m > n$.

   (a) Let w^* be one minimizer of the loss function L above. Give an infinite set of minimizers of the loss function.

   (b) What property defines the minimizer given by the representer theorem (in terms of X)?