Math Review Notes—Statistical Learning

Gregory Faletto

Contents

Last updated October 2, 2019

1 Statistical Learning

These notes are based on my notes from Math 547 at USC taught by Steven Heilman. I also borrowed from some other sources which I mention when I use them.

1.1 Math 547

Exercise 3.16: this inequality talks about the number of misclassifications, not the probability of misclassification under any distribution.

1.1.1 Perceptron Algorithm

Remark. Note that the run times only depends on the ℓ_2 norm of the solution loadings w and the ℓ_2 norm of the longest vector in the data set. That sounds good since it doesn't depend on the size of the data, but in the worse case θ can grow exponentially in the dimension of the data.

Also, note that the actual run time is at least linear in the size of the data, since on each iteration the algorithm checks some multiple of m points.

1.1.2 Mercer's Theorem

How is this an infinite-dimensional version of the Exercise? Let M be a $k \times k$ real symmetric matrix. By the Spectral Theorem, there exists an orthogonal Q and a diagonal D such that $M = Q^T DQ$. For all $1 \leq p \leq k$, let λ_p denote the pth diagonal entry of D. Let $\psi_i^p \in \mathbb{R}^k$ denote the ith row of Q. Then

$$m_{ij} = \sum_{p=1}^{k} \lambda_p \psi_i^{(p)} \psi_j^{(p)}, \qquad \forall 1 \leq i, j \leq k.$$

Also, $m(x, y)$ is called a **kernel.**

How to get ψ from $m(x, y)$? Definte

$$\ell_2 := \{\}$$

Note: if we could write the algorithm in terms of $m(x, y)$, we don't need to specify this embedding ϕ at all.

1.2 Norms

Proposition 1. Let $C \subseteq \mathbb{R}^n$ be a symmetric (if $c \in C$ then $-c \in C$) convex set containing 0. Define for all $x \in \mathbb{R}^n$,

$$\|x\|_C := \inf \left\{ t > 0 : \frac{x}{t} \in C \right\}$$

Then $\|x\|_C$ is a norm.