CMS 165 Foundations of Machine Learning Homework 1

In binary classification, the soft-margin SVM learning objective is:

$$\underset{w,b,\xi}{\operatorname{argmin}} \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\xi_i \tag{1}$$

$$\text{s.t.} \quad \forall i: \; y_i(w^T x_i - b) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \tag{2}$$

where the supervised training set is $S = \{(x_i, y_i)\}_{i=1}^N$ with $x_i \in \Re^D$ and $y_i \in \{-1, +1\}$, and $C \geq 0$ is a hyperparameter.

The hard-margin SVM is:

$$\underset{w,b}{\operatorname{argmin}} \frac{1}{2}\|w\|_2^2 \tag{3}$$

$$\text{s.t.} \quad \forall i: \; y_i(w^T x_i - b) \geq 1, \xi_i \geq 0 \tag{4}$$

Question 1. The soft-margin SVM problem is a constrained optimization problem with constraints specified in (2). Typically, in supervised learning, the learning objective we most commonly studied is unconstrained, e.g.:

$$\underset{w,}{\operatorname{argmin}} \frac{1}{2}\|w\|_2^2 + \frac{C}{N}\sum_{i=1}^{N}\ell(w, b, x_i, y_i), \tag{5}$$

where $\ell(w, b, x_i, y_i)$ is a convex loss function that measures the mismatch between $w^T x_i$ and y_i. Define ℓ such that (5) is equivalent to solving (1) & (2). (Hint: this is the hinge loss.)

$$\ell(x, b, y, w) = \max\left\{0, 1 - y(w^T x - b)\right\}.$$

Question 2. What happens in the soft-margin SVM problem as C grows from 0?

Question 3.

Derive the bias-variance decomposition for the squared error loss function. That is, prove that for a model f trained on a dataset S to predict a target $y(x)$ for each x, the following relation holds:

$$\mathbb{E}_S[E_{\text{out}}(f_S)] = \mathbb{E}_x[\text{Bias}(x) + \text{Var}(x)]$$

given the following definitions:

$$\begin{aligned} F(x) &= \mathbb{E}_S[f_S(x)] \\ E_{\text{out}}(f_S) &= \mathbb{E}_x[(f_S(x) - y(x))^2] \\ \text{Bias}(x) &= (F(x) - y(x))^2 \\ \text{Var}(x) &= \mathbb{E}_S[(f_S(x) - F(x))^2] \end{aligned}$$

Question 4. Let A be an $n \times n$ real symmetric matrix. Prove that the following two statements are equivalent: 1) all eigenvalues of A are greater than or equal to zero 2) for all vectors $x \in \mathbb{R}^n$, $x'Ax$ is greater than or equal to zero.

Recall that a naive Bayes model can be represented as:

$$P(x,y) = P(y)\prod_{d=1}^{D} P(x^d|y), \tag{6}$$

where x^d denotes the d-th feature entry of feature vector x. For simplicity, assume all x and y are binary.

In supervised training, the goal is to estimate the probability tables in (6) to optimize:

$$\operatorname{argmax} \prod_{(x_i,y_i)\in S} P(x_i,y_i) \equiv \operatorname{argmin} \sum_{(x_i,y_i)\in S} -\log P(x_i,y_i), \tag{7}$$

where $S = \{(x_i,y_i)\}_{i=1}^{N}$ is the supervised training set.

Question 5: Derive the maximum likelihood solution for supervised learning of naive Bayes, i.e., derive the solution to (7).

Question 6: What is the most common way to regularize when training naive Bayes models? Can you give a practical interpretation of it?

Question 7: Why is naive Bayes considered a generative model? How can one use generative models?

A simple HMM can be written as:

$$P(x,y) = P(y_0)\prod_{t=1}^{T} P(y_t|y_{t-1})P(x_t|y_t), \tag{8}$$

where y_0 denotes a special start state.

Question 8: Compare and contrast (6) with (8). Is there a unified framework that subsumes both?

Question 9: Let A and B be two $n \times n$ matrices. Show that the rank of AB is at most the minimum of the rank of A and the rank of B. Also show that if A, B are both non-singular than so is AB.

Question 10: Let A be an $n \times n$ matrix. Show that the non-zero singular values of A are the square-roots of the non-zero eigenvalues of AA'.