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.)

Solution:

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

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

Solution: When $C = 0$, then the soft-margin SVM will return $w = 0$ because the training error is not considered in the optimization. As C grows, the learning objective will increasingly favor smaller training loss over smaller model complexity (as measured in $\|w\|$).

Question 3.

Derive the bias-variance decomposition for the squared error loss function. That is, prove that for a model f_S 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:

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

Solution:
We have

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

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.

Solution: First we show that the first statement implies the second. Since A is real symmetric, it is orthogonally diagonalizable, i.e. there exists a basis such that A is a diagonal matrix in this basis, and furthermore the diagonal elements of A are its eigenvalues $\lambda_1 \dots \lambda_n$. Writing A in this basis, we see that $x'Ax = \sum_{i=1}^n \lambda_i x_i^2$ which is clearly non-negative since by assumption each λ_i was non-negative. Now we show that the second statement implies the first. Suppose by way of contradiction that there existed some eigenvalue λ which was negative. Let x be an (non-zero) eigenvector corresponding to λ. Then $x'Ax = \lambda\|x\|^2 < 0$, contradiction.

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:

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

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).

Solution: Define $P(y=1) = e^{b_1}/(e^{b_0}+e^{b_1})$, and $P(x=1|y=1) = e^{o_{11}}/(e^{o_{01}}+e^{o_{11}})$. Other probability quantities can be defined analogously. We can write the derivative w.r.t. b_1 as:

$$\begin{aligned}
\frac{\partial}{\partial b_1} \sum_{(x_i,y_i)\in S} -\log P(x_i,y_i) &= \sum_{(x_i,y_i)\in S} \frac{\partial}{\partial b_1} -\log P(x_i,y_i) \\
&= \sum_{(x_i,y_i)\in S:y_i=1} \frac{\partial}{\partial b_1} -\log \frac{e^{b_1}}{e^{b_0}+e^{b_1}} + \sum_{(x_i,y_i)\in S:y_i=0} \frac{\partial}{\partial b_1} -\log \frac{e^{b_0}}{e^{b_0}+e^{b_1}} \\
&= \sum_{(x_i,y_i)\in S:y_i=1} \frac{\partial}{\partial b_1} -b_1 + \log(e^{b_0}+e^{b_1}) + \sum_{(x_i,y_i)\in S:y_i=0} \frac{\partial}{\partial b_1} -b_0 + \log(e^{b_0}+e^{b_1}) \\
&= -N_{[y_i=1]} + N\frac{e^{b_1}}{e^{b_0}+e^{b_1}} \\
&= -N_{[y_i=1]} + NP(y=1),
\end{aligned}$$

where $N_{[y_i=1]}$ denotes the number of data points with $y_i = 1$. Setting the above derivative to 0 gives us $P(y=1) = N_{[y_i=1]}/N$, i.e., the fraction of training data with $y_i = 1$. Other entries can be derived analogously.

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

Solution: The most common way to regularize is to use pseudo counts. In the above solution that would be: $P(y=1) = (N_{[y_i=1]}+\lambda)/(N+2\lambda)$, where $lambda$ is the regularization strength. One can interpret this as hallucinating λ positive and λ negative data points.

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

Solution: NB is considered a generative model because it is a probabilistic model that can generative the test distribution (assuming, the test distribution was generated by some NB model). One can use generative models to sample complete (x, y) data points or any subset of features of the data points (via marginalization).

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?

Solution: In NB there is only one y and set of x. In HMM, there is a sequence of y's and x's, although each x only has a single feature rather than D features. In HMM, the x is often over many categories rather than binary as is often in NB. A unifed framework would be

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

which allows for multiple emissions at each token in the sequence (like in NB).

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.

Solution: By definition, rank(AB) = dim(range(AB)). Clearly, range$(AB) \subset$ range(A) and thus rank$(AB) \leq$ rank(A). Let $v_1, \ldots, v_k$ denote basis vectors for the subspace range(B) where $k =$ rank$(B) \leq n$; thus, range$(AB) = span(Av_1, \ldots, Av_k)$. Hence, rank$(AB) \leq k =$ rank(B). Therefore rank$(AB) \leq \min($rank$(A),$ rank$(B))$.

Since A, B are invertible, $B^{-1}A^{-1}AB = I$ and $ABB^{-1}A^{-1} = I$. Thus $(AB)^{-1} = B^{-1}A^{-1}$

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'.

Solution: Let $A = U\Sigma V'$ be the singular value decomposition of A, where U and V are unitary matrices and Σ is diagonal matrix with singular values. Then, $AA' = U\Sigma V'V\Sigma U' = U\Sigma^2 U'$. This is the eigendecomposition of AA'. Thus, the statement follows.