CS 189 Spring 2022 | Introduction to Machine Learning Jonathan Shewchuk

Exam Prep 1 Solutions

This exam-prep discussion section covers Bayesian decision theory and maximum likelihood estimation. In order, the questions were taken from the Spring offerings in 2016, 2016, 2017, 2019, and 2017.

1 Multiple Choice

[3 pts] In the usual formulation of soft-margin SVMs, each training sample has a slack variable $\xi_i \geq 0$ and we impose a regularization cost $C \sum_i \xi_i$. Consider an alternative formulation where we impose the additional constraints $\xi_i = \xi_j$ for all i, j. How does the minimum objective value $|\mathbf{w}|^2 + C \sum_i \xi_i$ obtained by the new method compare to the one obtained by the original soft-margin SVM?

- ○ They are always equal.
- ○ Original SVM minimum $\geq$ new minimum.
- ● New minimum $\geq$ original SVM minimum.
- ○ New minimum is sometimes larger and sometimes smaller.

(f) [3 pts] Which of the following holds true when running an SVM algorithm?

- ● Increasing or decreasing α value only allows the decision boundary to translate.
- ○ Decision boundary rotates if we change the constraint to $w^T x + \alpha \geq 3$.
- ● Given n-dimensional points, the SVM algorithm finds a hyperplane passing through the origin in the $(n+1)$-dimensional space that separates the points by their class.
- ● The set of weights that fulfill the constraints of the SVM algorithm is convex.

(b) [4 pts] Which of the following changes would commonly cause an SVM's margin $1/\|w\|$ to shrink?

- ● A: Soft margin SVM: increasing the value of C
- ○ C: Soft margin SVM: decreasing the value of C
- ● B: Hard margin SVM: adding a sample point that violates the margin
- ○ D: Hard margin SVM: adding a new feature to each sample point

The greater the value of C is, the higher the penalty for violating the margin. The soft margin shrinks to compensate.

If you add a sample point that violates the margin, a hard margin always shrinks.

If you add a feature, the old solution can still be used (by setting the weight associated with the new feature to zero). Although the new feature might enable a new solution with a wider margin, the optimal solution can't be worse than the old solution.

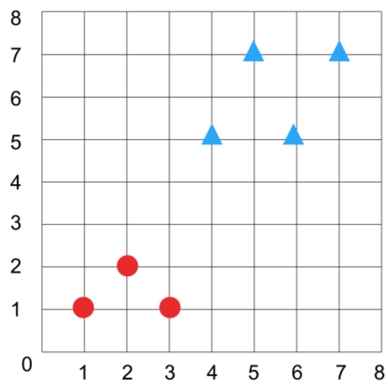
2 Free Response

Q2. [20 pts] Hard-Margin Support Vector Machines

Recall that a **maximum margin classifier**, also known as a hard-margin support vector machine (SVM), takes n training points $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ with labels $y_1, y_2, \ldots, y_n \in \{+1, -1\}$, and finds parameters $w \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}$ that satisfy a certain objective function subject to the constraints

$$y_i(X_i \cdot w + \alpha) \geq 1, \quad \forall i \in \{1, \ldots, n\}.$$

For parts (a) and (b), consider the following training points. Circles are classified as positive examples with label $+1$ and triangles are classified as negative examples with label -1.



(a) [3 pts] Which points are the support vectors? Write it as $\begin{bmatrix}\text{horizontal}\\ \text{vertical}\end{bmatrix}$. E.g., the bottom right circle is $\begin{bmatrix}3\\1\end{bmatrix}$.

The support vectors are the points $\begin{bmatrix}2\\2\end{bmatrix}$ and $\begin{bmatrix}4\\5\end{bmatrix}$.

(b) [4 pts] If we add the sample point $x = \begin{bmatrix}5\\1\end{bmatrix}$ with label -1 (triangle) to the training set, which points are the support vectors?

The support vectors are the points $\begin{bmatrix}3\\1\end{bmatrix}$, $\begin{bmatrix}4\\5\end{bmatrix}$, and $\begin{bmatrix}5\\1\end{bmatrix}$.

For parts (c)–(f), forget about the figure above, but assume that there is at least one sample point in each class and that the sample points are linearly separable.

(c) [2 pts] Describe the geometric relationship between w and the decision boundary.

The weight vector w (called the *normal vector*) is orthogonal to the decision boundary.

(d) [2 pts] Describe the relationship between w and the margin. (For the purposes of this question, the margin is just a number.)

The margin (the distance from the decision boundary to the nearest sample point) is $1/\|w\|$.

(e) [4 pts] Knowing what you know about the hard-margin SVM objective function, explain why for the optimal (w, α), there must be at least one sample point for which $X_i \cdot w + \alpha = 1$ and one sample point for which $X_i \cdot w + \alpha = -1$.

The objective is to minimize $\|w\|^2$ (or equivalently, $\|w\|$). If every sample point has $y_i(X_i \cdot w + \alpha) > 1$, we can simply scale w to make it smaller until there is a point such that $y_i(X_i \cdot w + \alpha) = 1$, thereby improving the "solution."

If we have a positive sample point for which $X_i \cdot w + \alpha = 1$ but every negative sample point has $X_i \cdot w + \alpha < -1$, we can make α a little greater so that every sample point has $y_i(X_i \cdot w + \alpha) > 1$. Then we can shrink w some more. So any such "solution" cannot be optimal. (The symmetric argument applies if a negative sample point touches the slab but not positive sample point does.)

(f) [5 pts] If we add new features to the sample points (while retaining all the original features), can the optimal $\|w_{\text{new}}\|$ in the enlarged SVM be greater than the optimal $\|w_{\text{old}}\|$ in the original SVM? Can it be smaller? Can it be the same? Explain why! (Most of the points will be for your explanation.)

It can be smaller, or it can be the same, but it cannot be greater.

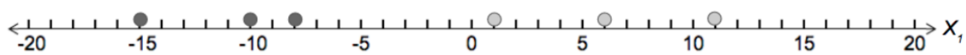
If w_{old} and α are an optimal solution of the original SVM, when we add features we can create a w_{new} that has the same values as w_{old}, with zeros added for the new features. Then w_{new} and α satisfy all the constraints of the enlarged SVM. These might not be the optimal solution, but the optimal solution of the enlarged SVM cannot have $\|w_{\text{new}}\|$ greater than $\|w_{\text{old}}\|$.

$\|w_{\text{new}}\|$ can be smaller, because the new features can put an arbitrarily large amount of space between the classes, making the margin arbitrarily large.

$\|w_{\text{new}}\|$ will be the same as $\|w_{\text{old}}\|$ if the new features are all zeros in all the sample points.

Q2. [10 pts] Comparing Classification Algorithms

Find the decision boundary given by the following algorithms. Provide a range of values if the algorithm allows for multiple feasible decision boundaries. If there exists no feasible decision boundary, state "None."



(a) [1 pt] Perceptron: X_1 = ________________

(b) [2 pts] Hard-Margin SVM: X_1 = ________________

(c) [2 pts] Linear Discriminant Analysis: X_1 = ________________

Perceptron: [-8, 1]
Hard-Margin SVM: -3.5
Linear Discriminant Analysis: -2.5