CS 475 Machine Learning: Homework 3
Supervised Classifiers 2
Due: Friday October 16, 2015, 11:59pm
100 Points Total Version 1.1

Li-Yi Lin / llin34@jhu.edu

1 Analytical (50 points)

The following problems consider a standard binary classification setting: we are given n observations with m features, $x_1, ..., x_n \in \mathbb{R}^m$.

1) Overfitting (8 points) SVMs using nonlinear kernels usually have two tuning parameters (regularization parameter C and kernel parameter γ), which are usually determined by cross validation.

(a) Suppose we use cross validation to determine the non-linear kernel parameter and slack variable for an SVM. We find that classifiers using parameters (c_1, γ_1) and (c_2, γ_2) achieve the same cross validation error, but (c_1, γ_1) leads to fewer support vectors than (c_2, γ_2). Explain which set of parameters should we choose for the final model?
Ans:
When two sets of parameters achieve the same cross validation error, we should choose the model with lower variance, which is more general (simple), for the final model. The reason is that, under the same training data, if one model has more support vectors, it means the model uses more features (higher dimension) for the kernel. And a model uses more features (higher dimension) for the kernel means it is more likely to be overfitting. Therefore, we should choose the model with fewer support vectors.

(b) The optimization problem of linear SVMs can be in either primal or dual form. As we know, the primal form has m parameters to learn, while the dual form has n parameters to learn. If $m \gg n$, is it true that the dual form reduces over-fitting since it has fewer parameters? Explain.
Ans:
The dual form will not necessarily reduces over-fitting problem even if $m >> n$. For the over-fitting concern, we should not only look at the number of parameters, but also look at the dimension of a kernel that the model uses. When the dimension of a kernel is higher, that means it will make the model more likely to be over-fitting to the training data. The impact on the over-fitting from the higher dimensional kernel might be larger than that from the large number of parameters. Thus, the number of parameters is not the only factor that affects the degree of over-fitting.

2) Hinge Loss (12 points) Linear SVMs can be formulated in an unconstrained optimization problem

$$\min_w \sum_{i=1}^{n} H(y_i(w^T x_i)) + \lambda \|w\|_2^2, \tag{1}$$

where λ is the regularization parameter and $H(a) = \max(1-a, 0)$ is the well known hinge loss function. The hinge loss function can be viewed as a convex surrogate of the 0/1 loss function $I(a \leq 0)$.

(a) Prove that $H(a)$ is a convex function of a.

Ans:

If $H(a)$ function is convex, then it must satisfy

$$H(ta_1 + (1-t)a_2) \leq tH(a_1) + (1-t)H(a_2)$$

$$\text{where } 0 < t < 1$$

We substitute $H(a) = \max(1-a, 0)$ to the above equation, then we have to prove the following equation

$$\max(ta_1 + (1-t)a_2, 0) \leq t\max(a_1) + (1-t)\max(a_2)$$

Since the value of $H(a) = \max(1-a, 0)$ function will be 0 if $a \geq 1$ and $1-a$ if $a < 1$, we will discussion three situations: (1) $1 \leq a_1 < a_2$; (2) $a_1 < a_2 < 1$; (3) $a_1 < 1 < a_2$.
(1) when $1 \leq a_1 < a_2 => 1 < ta_1 + (1-t)a_2$

$$H(ta_1 + (1-t)a_2) = 0$$

$$H(a_1) + H(a_2) = 0 + 0 = 0$$

Since $0 = 0$, we have proved that

$$H(ta_1 + (1-t)a_2) \leq tH(a_1) + (1-t)H(a_2)$$

(2) when $a_1 < a_2 < 1 => ta_1 + (1-t)a_2 < 1$

$$H(ta_1 + (1-t)a_2) = 1 - (ta_1 + (1-t)a_2) = 1 - ta_1 - a_2 + ta_2$$
$$tH(a_1) + (1-t)H(a_2) = t(1-a_1) + (1-t)(1-a_2) = 1 - ta_1 - a_2 + ta_2$$

Since $1 - ta_1 - a_2 + ta_2 = 1 - ta_1 - a_2 + ta_2$, we have proved that

$$H(ta_1 + (1-t)a_2) \leq tH(a_1) + (1-t)H(a_2)$$

(3) if $a_1 < 1 < a_2$:
(3.1) when $ta_1 + (1-t)a_2 \geq 1 => 1 - (ta_1 + (1-t)a_2) \leq 0$

$$H(ta_1 + (1-t)a_2) = \max(1 - (ta_1 + (1-t)a_2), 0) = 0$$

$$tH(a_1) + (1-t)H(a_2) = t(1-a_1) + 0 = t(1-a_1)$$

Since $t(1-a_1) > 0$, we have proved that

$$H(ta_1 + (1-t)a_2) \leq tH(a_1) + (1-t)H(a_2)$$

(3.2) when $ta_1 + (1-t)a_2 < 1 => 1 - (ta_1 + (1-t)a_2) > 0$

$$H(ta_1 + (1-t)a_2) = \max(1 - (ta_1 + (1-t)a_2), 0) = 1 - (ta_1 + (1-t)a_2)$$

$$tH(a_1) + (1-t)H(a_2) = t(1-a_1) + (1-t)0 = t(1-a_1)$$

Since

$$\begin{aligned} &tH(a_1) + (1-t)H(a_2) - H(ta_1 + (1-t)a_2) \\ &= t(1-a_1) - (1 - (ta_1 + (1-t)a_2)) \\ &= t - ta_1 - 1 + ta_1 + a_2 - ta_2 \\ &= t(1-a_2) - (1-a_2) \\ &= (t-1)(1-a_2) > 0 \end{aligned}$$

we have $tH(a_1) + (1-t)H(a_2) - H(ta_1 + (1-t)a_2) > 0$. Therefore, we have proved that $H(ta_1 + (1-t)a_2) \leq tH(a_1) + (1-t)H(a_2)$. Thus, $H(a)$ is a convex function.

(b) The function $L(a) = \max(-a, 0)$ can also approximate the 0/1 loss function. What is the disadvantage of using this function instead?
Ans:

(c) If $H'(a) = \max(0.5 - a, 0)$, show that there exists λ' such that (2) is equivalent to (1). Hint: think about the geometric interpretation of hinge loss.

$$\min_w \sum_{i=1}^{n} H'(y_i(w^T x_i)) + \lambda' \|w\|_2^2. \tag{2}$$

Ans:
Since the meaning will not change even if we scale the w, we can scale w by multiplying it with 0.5. Then we get the following objective function:

$$\min_w \sum_{i=1}^{n} H'(y_i(0.5w^T x_i)) + \lambda' \|0.5w\|_2^2$$

$$= \min_w \sum_{i=1}^{n} \max(0.5 - y_i(0.5w^T x_i)) + \lambda' \|0.5w\|_2^2$$

$$= \min_w \sum_{i=1}^{n} 0.5 \max(1 - y_i(w^T x_i)) + 0.25\lambda' \|w\|_2^2$$

$$= 0.5 \min_w \sum_{i=1}^{n} \max(1 - y_i(w^T x_i)) + 0.5\lambda' \|w\|_2^2$$

When $0.5\lambda' = \lambda => \lambda' = 2\lambda$, we have found a λ' such that (2) is equivalent to (1) because the solution of w will be the same.

3) Kernel Trick (10 points) The kernel trick extends SVMs to handle with nonlinear data sets. However, an improper use of a kernel function can cause serious over-fitting. Consider the following kernels.

(a) Polynomial kernel: $K(x, x') = (1 + (x^T x'))^d$, where $d \in \mathbb{N}$. Does increasing d make over-fitting more or less likely?
Ans:
The polynomial kernel $K(x, x') = (1+(x^T x'))^d$ can be written as $\sum_{n=0}^{d} \binom{d}{n} 1^n (x^T x')^{d-n}$. As we can see, when d increases, the kernel will have more dimensions and more dimensions will make the model over-fitting more likely. Therefore, increasing d will make over-fitting more likely.

(b) Gaussian kernel: $K(x, x') = \exp(-\|x - x'\|^2/2\sigma^2)$, where $\sigma > 0$. Does increasing σ make over-fitting more or less likely?
Ans:

We say K is a kernel function, if there exists some transformation $\phi : \mathbb{R}^m \to \mathbb{R}^{m'}$ such that $K(x_i, x_{i'}) = \langle \phi(x_i), \phi(x_{i'}) \rangle$.

(c) Let K_1 and K_2 be two kernel functions. Prove that $K(x_i, x_{i'}) = K_1(x_i, x_{i'}) + K_2(x_i, x_{i'})$ is also a kernel function.
Ans:

$$\begin{aligned} K_1(x_i, x_{i'}) + K_2(x_i, x_{i'}) &= \langle \phi_1(x_i), \phi_1(x_{i'}) \rangle + \langle \phi_2(x_i), \phi_2(x_{i'}) \rangle \\ &= \sum_{n=1}^{m_1} \phi_{1n}(x_i)\phi_{1n}(x_{i'}) + \sum_{n=1}^{m_2} \phi_{2n}(x_i)\phi_{2n}(x_{i'}) \\ &= [\phi_{1n}(x_i), \phi_{2n}(x_i)]^T [\phi_{1n}(x_{i'}), \phi_{2n}(x_{i'})] \end{aligned}$$

We let $\phi_3(x_i) = [\phi_{1n}(x_i), \phi_{2n}(x_i)]$ and $\phi_3(x_{i'}) = [\phi_{1n}(x_{i'}), \phi_{2n}(x_{i'})]$. In addition, we let $K(x_i, x_{i'}) = \langle \phi_3(x_i), \phi_3(x_{i'}) \rangle$. Then we have

$$[\phi_{1n}(x_i), \phi_{2n}(x_i)]^T [\phi_{1n}(x_{i'}), \phi_{2n}(x_{i'})] = \langle \phi_3(x_i), \phi_3(x_{i'}) \rangle$$

Hence, we have proved that

$$K(x_i, x_{i'}) = K_1(x_i, x_{i'}) + K_2(x_i, x_{i'})$$

4) Prediction using Kernel (8 points) One of the differences between primal linear SVMs and dual kernel SVMs concerns computational complexity at prediction time.

(a) What is the computational complexity of prediction of a primal linear SVM in terms of the numbers of the training samples n and features m?
Ans:

(b) What is the computational complexity of prediction of a dual kernel SVM in terms of the numbers of the training samples n, features m, and support vectors s?
Ans:

5) Stochastic Gradient Algorithm (12 points) The stochastic gradient algorithm is a very powerful optimization tool to solve large-scale machine learning problems. Instead of computing the gradient over the entire data set before making an update, the stochastic gradient algorithm computes the gradient over a single sample, then updates the parameters. By passing over the entire data set of n samples in this fashion we can converge to the optimal parameters.

A single iteration of stochastic gradient considers a single example. While it takes many more iterations, each iteration is much faster, both in terms of memory and computation.

Consider a ridge regression problem with n samples:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_2^2. \tag{3}$$

In each iteration, instead of using only one example, we randomly choose k out of n samples and obtain $(x_{1'}, y_{1'}), ..., (x_{k'}, y_{k'})$.

(a) What is the computational complexity of computing the mini-batch stochastic gradient or gradient at each iteration (using k and n samples respectively)?

(b) What are the advantages/disadvantages of increasing k in terms of computational complexity (using k and n samples respectively)? What is traded-off by increasing/decreasing k?

(c) Give one advantage and one disadvantage of using stochastic gradient descent with $k = 1$ for a possibly nonconvex optimization problem, ignoring computational considerations. Explain.

2 What to Submit

In each assignment you will submit two things.

1. **Code:** Your code as a zip file named `library.zip`. **You must submit source code (.java files)**. We will run your code using the exact command lines described above, so make sure it works ahead of time. Remember to submit all of the source code, including what we have provided to you.

2. **Writeup:** Your writeup as a **PDF file** (compiled from latex) containing answers to the analytical questions asked in the assignment. Make sure to include your name in the writeup PDF and use the provided latex template for your answers.

Make sure you name each of the files exactly as specified (library.zip and writeup.pdf).

To submit your assignment, visit the "Homework" section of the website (http://www.cs475.org/.)

3 Questions?

Remember to submit questions about the assignment to the appropriate group on the class discussion board: http://bb.cs475.org.