Homework Assignment 2
Loss Functions and Support Vector Machines

Lecturer: Kyunghyun Cho

February 28, 2017

1. After replacing the label set from $\{0,1\}$ to $\{-1,1\}$, we introduced the log loss

$$D_{\log}(y,\mathbf{x};M) = \frac{1}{\log 2}\log(1+\exp(-s(y,\mathbf{x};M))),$$

as an alternative to the logistic regression distance function above. Show that these two are equivalent up to a constant multiplication for logistic regression.

SOLUTION: Let us say that the log loss for $\{0,1\}$ is $f(y) = -(y^*\log M(\mathbf{x}) + (1-y^*)\log(1-M(\mathbf{x})))$ and the log loss for $\{-1,1\}$ is $g(y) = \frac{1}{\log 2}\log(1+e^{-yw^T\mathbf{x}}))$. We essentially want to prove that $f(0) = g(-1)$ and $f(1) = g(1)$.
$f(0) = -\log(1-\frac{1}{1+e^{-w^T\mathbf{x}}}) = -\log(\frac{e^{-w^T\mathbf{x}}}{1+e^{-w^T\mathbf{x}}}) = \log(\frac{1+e^{-w^T\mathbf{x}}}{e^{-w^T\mathbf{x}}}) = \log(1+e^{w^T\mathbf{x}})$.
$g(-1) = \frac{1}{\log 2}\log(1+e^{w^T\mathbf{x}})$
$f(0) = \frac{1}{\log 2}g(-1)$

$f(1) = -\log(\frac{1}{1+e^{-w^T\mathbf{x}}}) = \log(1+e^{-w^T\mathbf{x}})$
$g(1) = \frac{1}{\log 2}\log(1+e^{-w^T\mathbf{x}})$
$f(1) = \frac{1}{\log 2}g(1)$

Therefore, both equations are equivalent up to $\frac{1}{\log 2}$.

2. Unlike the log loss, the hinge loss, defined below, is not differentiable everywhere:

$$D_{\text{hinge}}(y,\mathbf{x};M) = \max(0,1-s(y,\mathbf{x};M)).$$

Does it mean that we cannot use a gradient-based optimization algorithm for finding a solution that minimizes the hinge loss? If not, what can we do about it?

SOLUTION: For points > 1, the derivative is 0. For points < 1, the hinge function is equal to $1-y\mathbf{w}^\top\tilde{\mathbf{x}}$. This case has a trivial derivative of $-y\tilde{\mathbf{x}}$. The only part which we cannot find the derivative is at the point where $s(y,\mathbf{x};M) = 1$. Instead, we can simply

arbitrarily define the derivative to be 0 at that point because as far as we are concerned, a value should be considered correctly classified if it is equal to exactly 1. Anything greater is a correct classification and values slightly below are barely classified.

3. When the distances to the nearest positive and negative examples are defined as d^+ and d^-, the margin is

$$\gamma = \frac{1}{2}(d^+ + d^-).$$

Show that minimizing the norm of the weight vector of a support vector machine is equivalent to maximizing the margin.

SOLUTION: We know that the margin is the distance of the closest examples to the decision boundary. This value should be maximized in order to increase the probability of correctly classifying our test data and to minimize our empirical cost. when substituting the values into the above equation, we get:
$\gamma = \frac{1}{2}(\frac{w^T\mathbf{x}^+}{||w||} + \frac{w^T\mathbf{x}^-}{||w||}) = \frac{w^T\mathbf{x}+b}{||w||}$
The best way to set this up is to normalize the support vectors such that they are a distance of 1 away from the decision boundary. This gives us $\gamma = \frac{2}{||w||}$. This shows that the margin is inversely proportional to the norm of the weight vector, thus demonstrating that minimizing this value is equivalent to maximizing the margin.

4. PROGRAMMING ASSIGNMENT