Loss Functions and Support Vector Machines

Neal Moorthy

2/15/18

1. In the lecture notes the distance function is given as $D(y, x; M) = -(ylog(M(x)) + (1 - y)log(1 - M(x)))$. The distance function with log loss is $D_{log}(y, x; M) = \frac{1}{log2}log(1 + exp(-s(y, x; M)))$. There are two cases for y. In the original, equation $y \in \{0, 1\}$ to $\hat{y} \in \{1, 1\}$. In both equations when $y = 1$ the distance function is

$$\begin{aligned} D(1, x; M) &= -(log(M(x))) \\ &= -log(\frac{1}{1 + exp - w^T x}) \\ &= log(1 + exp - w^T x) \end{aligned} \tag{1}$$

meanwhile, logloss is

$$\begin{aligned} D_{log}(1, x; M) &= \frac{1}{log2}log(1 + exp(-s(1, x; M))) \\ &= \frac{1}{log2}log(1 + exp(-w^T x)) \end{aligned} \tag{2}$$

So the canstant is $\frac{1}{log2}$ for the positive case. What about in the negative case?

$$\begin{aligned} D(0, x; M) &= -(1 - log(M(x))) \\ &= -log(1 - \frac{1}{1 + exp - w^T x}) \\ &= -log(\frac{1}{1 + expw^T x}) \\ &= log(1 + expw^T x) \end{aligned} \tag{3}$$

meanwhile, logloss is

$$\begin{aligned} D_{log}(-1, x; M) &= \frac{1}{log2} log(1 + exp(-s(-1, x; M))) \\ &= \frac{1}{log2} log(1 + exp(w^T x)) \end{aligned} \tag{4}$$

So in the end both functions are the same up to the constant $\frac{1}{log2}$ because we want our log-loss to be an upper bound on the 0-1 loss function. Without this constant then where $w = 0$ then D_{log} would be 0.

2. Hinge loss is only non-differentiable because of the hinge point. before that point we can still take the gradient and after that the gradient is 0. So if we consider the gradient of the hinge loss function to be the subgradient prior to the margin then we can use that to optimize.

3.

a. Pick i and t such that $min(\tilde{R}^i_{val_t})$

b. The generalization error reported is $\tilde{R}^i_{test_t} - \tilde{R}^i_{train_t}$.