Homework Assignment 1
Logistic Regression, Support Vector Machines

Kailin Zheng kz882

February 20, 2019

1 Empirical vs. expected cost (10 points)

When we use empirical cost to approximate the expected cost,

$$\mathbb{E}_{x\sim X}[\Delta(M^*(x), M, x)] + \lambda C(M) \approx \frac{1}{N}\Delta(M^*(x^n), M, x^n) + \lambda C(M) + \epsilon$$

is it okay to weigh each per-example cost equally? Given that we established that not every data x is equally likely, is taking the sum of all per-example costs and dividing by N reasonable? Should we weigh each per-example cost differently, depending on how likely each x is? Justify your answer.

Answer: Since we established that not every data x is equally likely, taking the sum of all per-example costs and dividing by N is not reasonable. That is to say, in the video-intrusion example in class notes, it would not be reasonable to sum 0+10+100 and divide by 3.

Ideally, it would be the best to weigh each per-example cost according to the probability that event (or x) happens. If we know p(no thief), p(thief with detection), p(thief without detection), we will be able to weigh each x depending on how likely each x is.

However, in reality, we cannot know the probability for each per-example cost. But we can have a hypothesis set, a finite set of samples. This approximation method, that is approximation based on a finite set of samples from a probability distribution, is called a Monte Carlo method.

The idea is, if we put random samples into the sample set, the hypothesis set should to some extent reflect the probability distribution in the reality. Thus, when we divide by N for every per-example cost, it should be consistent with the ("average") cost in reality.

2 Logistic loss gradient (10 points)

The distance function of logistic regression was defined as

$$\Delta(y^*, w, x) = -(y^* log M(x) + (1 - y^*) log(1 - M(x)))$$

Derive its gradient with respect to the weight vector w step-by-step.

Answer: Gradient with respect to weight vector w is:
$\frac{\partial \Delta(y^*,w,x)}{\partial w} = -y^* \frac{\partial log M(x)}{\partial w} - (1-y^*)\frac{\partial log(1-M(x))}{\partial w}$.

Note that $M(x) = \sigma(w^T\tilde{x})$, $\frac{\partial log M(x)}{\partial w} = \frac{\frac{\partial M(x)}{\partial w}}{M(x)}$,
$\frac{\partial log(1-M(x))}{\partial w} = -\frac{\frac{\partial M(x)}{\partial w}}{1-M(x)}$.

The derivative of Sigmoid function $\sigma(a) = \frac{1}{1+exp(-a)}$ is
$\frac{d\sigma(a)}{da} = \sigma(a)(1-\sigma(a))$.

So $\frac{\partial M(x)}{\partial w} = M(x)(1-M(x))\tilde{x}$.

Insert $\frac{\partial M(x)}{\partial w}$ to the logs, we have:
$\frac{\partial log M(x)}{\partial w} = \frac{M(x)(1-M(x))\tilde{x}}{M(x)} = (1-M(x))\tilde{x}$,
$\frac{\partial log(1-M(x))}{\partial w} = -\frac{M(x)(1-M(x))\tilde{x}}{1-M(x)} = -M(x)\tilde{x}$.

So the gradient $\frac{\partial \Delta(y^*,w,x)}{\partial w} = -y^*(1-M(x))\tilde{x} - (1-y^*)(-M(x)\tilde{x})$
$= -(y^* - M(x))\tilde{x}$.

3 Hinge loss gradients (10 points)

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

$$\Delta hinge(y,x;M) = max(0, 1-s(y,x;M))$$

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?

Answer: Despite the fact that the hinge loss is not differentiable everywhere, we can still use a gradient-based optimization algorithm where the hinge loss IS differentiable.

Note that we have $s(y,x;M) = yw^T\tilde{x}$.

- When $1-s(y,x;M) > 0$, we have $\Delta hinge(y,x;M) = max(0, 1-s(y,x;M)) = 1-s(y,x;M)$.

  $\nabla w \Delta hinge(y,x;M) = \frac{\partial}{\partial w}(1-s(y,x;M)) = \frac{\partial}{\partial w}(1-yw^T\tilde{x}) = -y\tilde{x}$
- When $1-s(y,x;M) < 0$, we have $\Delta hinge = 0$, $\nabla w \Delta hinge(y,x;M) = 0$.
- When $1-s(y,x;M) < 0$, $\Delta hinge = 0$ is not differentiable.

To summarize, where hinge is differentiable, we have gradient:

$$\frac{\partial \Delta hinge(y,x;M)}{\partial w} = \begin{cases} -y\tilde{x} & \text{if s(y,x;M)} < 1 \\ 0 & \text{if s(y,x;M)} > 1 \end{cases}$$