CS 475 Machine Learning: Homework 1
Learning Foundations

Due: Wednesday February 12, 2014, 12:00pm

50 Points Total Version 1.2

Yiran Zhang

1 Analytical (15 Points)

In addition to completing the analytical questions, your assignment for this homework is to learn Latex. All homework writeups must be PDFs compiled from Latex. Why learn latex?

1. It is incredibly useful for writing mathematical expressions.
2. It makes references simple.
3. Many academic papers are written in latex.

The list goes on. Additionally, it makes your assignments much easier to read than if you try to scan them in or complete them in Word.

We realize learning latex can be daunting. Fear not. There are many tutorials on the Web to help you learn. We recommend using pdflatex. It's available for nearly every operating system. Additionally, we have provided you with the tex source for this PDF, which means you can start your writeup by erasing much of the content of this writeup and filling in your answers. You can even copy and paste the few mathematical expressions in this assignment for your convenience. As the semester progresses, you'll no doubt become more familiar with latex, and even begin to appreciate using it.

Be sure to check out this cool latex tool for finding symbols. It uses machine learning! `http://detexify.kirelabs.org/classify.html`

1 (3 points) A basket contains black and green grapes, 30% of the grapes are sweet, the rest are sour. 40% of the grapes that are sweet are black, 20% of the sour grapes are black. What is the probability that a black grape is sweet?

So from the question we know that:

1. $p(sweet) = 0.3 = 1 - p(sour)$
2. $p(black|sweet) = 0.4$
3. $p(black|sour) = 0.2$

And we are looking for $p(sweet|black)$

$$\begin{aligned} p(sweet|black) &= \frac{p(sweet, black)}{p(black)} \\ &= \frac{p(black|sweet)p(sweet)}{p(black, sweet) + p(black, sour)} \\ &= \frac{0.4\dot{0}.3}{0.4 \cdot 0.3 + (1 - 0.3) \cdot 0.2} \\ &= \frac{6}{13} = 0.46 \end{aligned}$$

2 (3 points) You are provided a computer program that produces random integers between 1 and 6, i.e. a die. The programmer advises you that the die results are not chosen IID. You are told that the die is biased. In order to determine its bias, you run the program for many trials, recording the number of times each number is returned. Suppose that out of n trials, there were m 1s. You are then asked to compute the probability that the next roll of the die will produce a 1. In terms of m and n, can you estimate the probability of a 1? If yes, what is it? If not, why not?

NO, this is because we cannot assume iid of the sampled data. Think about we have a dice that will give 6 with probability 1 in the following rolls if the first roll is not 1 and will give 1 if the first roll is 1. Then the sequence of observed data is very depend on the first data point. If it is not, then $p(6) = 1$ for all the following, then we will have the estimation that $p(1) = 0$ or $p(1) = 1$. But neither of these are nice estimation of the probability that are desired. To conclude, the iid assumption is essential if we want to do statistical estimation.

3 (4 points) For each of the following, state if the function is a valid loss function. If it is not, why not? Note that $\hat{y}$ is the predicted label and y is the correct label.

1. $\ell(y, \hat{y}) = y - \hat{y}$ NO, because the loss function can be both positive and negative. Consider that sequence $0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0$, if we use this loss function, then 0.5 will be a perfect hypothesis. But actually, there is quite a lot distance.

2. $\ell(y, \hat{y}) = (y - \hat{y})^2$ YES, this is a good and common loss function. It is even better than the loss function $y - \hat{y}$ since it is differentiable.

3. $\ell(y, \hat{y}) = |(y - \hat{y})|/\hat{y}$ NO, loss function may be both positive and negative, so the argument in (a) can also be used here. What is making things worse is that if the predicted label is 0, the loss function would blow up to infinity.

4 (5 points) Give an example of an optimal hypothesis, a finite hypothesis class that contains the optimal hypothesis, and an infinite class that does not contain the optimal hypothesis.

We are considering the problem that data are sampled from $y = x$ and we are trying to predict the mean. So the data should be like $(1, 1), (2, 2), (3.54, 3.54) \ldots$

Optimal hypothesis: $y = x$.

A finite hypothesis class: $y = kx, k$ is integer, $|k| <= 5$.

Infinite hypothesis class: $y = z \log x, z$is real.