Carnegie Mellon University, Machine Learning 10-701 February 2, 2015
http://alex.smola.org/teaching/10-701-15 due on February 9, 2015

Homework 3

START HERE: Instructions

- The homework is due at 9:00am on February 9, 2015. Anything that is received after that time will not be considered.
- Answers to every theory questions will be also submitted electronically on Autolab (PDF: Latex or handwritten and scanned). Make sure you prepare the answers to each question separately.
- Collaboration on solving the homework is allowed (after you have thought about the problems on your own). However, when you do collaborate, you should list your collaborators! You might also have gotten some inspiration from resources (books or online etc...). This might be OK only after you have tried to solve the problem, and couldn't. In such a case, you should cite your resources.
- If you do collaborate with someone or use a book or website, you are expected to write up your solution independently. That is, close the book and all of your notes before starting to write up your solution.
- Latex source of this homework: http://alex.smola.org/teaching/10-701-15/homework/hw2_latex.tar

1 Tail Bounds [Zhou; 50+bonus 20]

1.1 41 Shades of Blue [30 pts]

Before Google Instant, probably the most infamous example of Google's design-by-testing approach were the 41 shades of blue. Google's engineers apparently couldn't decide on two shades of blue for showing search results, so they tested 41 of them to see which attracted the most clicks.[1] They divided the users into 41 groups and tried to measure the user behaviors. Suppose that the click-through rate is in the order of 5%, with hard upper bound being 7%. Find out how long and how many subjects (i.e. users), would it be needed to carry out the experiment before concluding which blue is the best and up to which precision. Hint: if you need other variables to solve the problem, make an appropriate assumption for it.

Derive the bounds using two different strategies. How tight can you make your bounds? Do you need to test all 41 shades until the end?

Answer:

Assume clicks are modeled as a binomial distribution with the values 0 or 1, and that we wish to ensure a confidence bound δ of 0.05 within a bias ϵ of 0.002, we present two strategies for deriving the tail bounds:
Hoeffding Tail Bound:

$$m \leq -\frac{c^2 log(\frac{\delta}{2})}{2\epsilon^2} = -\frac{log(\frac{0.05}{2})}{2(0.002)^2} = 200257 \tag{1}$$

Chebyshev Tail Bound: If we assume a hard upper bound on click rate of 0.07, we then set a bound on σ^2=0.07(0.93) = 0.0651

$$m \leq \frac{\sigma^2}{\epsilon^2 \delta} = \frac{0.0651}{0.05 * (0.002)^2} = 325500 \tag{2}$$

Since there are 41 test groups, the number of ad impressions needed to be served before a winner can be selected with 95% confidence is 41*m, or approx. 13.35 million for the Chebyshev tail bound and 8.21 million for the Hoeffding tail bound. According to this link1, Google served 30 billion ads per day in 2012. If Google tracked 10 million instances of a certain link for testing purposes every day, it would only take 32.0 hours to complete the Chebyshev test and 19.7 hours to complete the Hoeffding test. However, there is a simple heuristic which can reduce the number of needed ad serves: if an empirical estimate of is

[1] http://www.fastcompany.com/1403230/googles-marissa-mayer-assaults-designers-data

periodically computed for all 41 testing groups, taking into account the current number of ads m which have been served, colors can be eliminated with a high degree of certainty if their upper confidence bound is lower than the lower confidence bound of the current leader. If we use the Chebyshev inequality as an example, at a given time point, compute the following:

$$\epsilon \leq \frac{\sigma^2}{m\delta} \tag{3}$$

for i in all the colors, if $\mu_{max} - \epsilon$ then we discard this color. Thus we can reduce the number the number of color we need to go through earlier.

1.2 Influence Factor of the Click Through Rate [20 pts]

Present a reason why the estimates of the click-through rate might be much better than what your bounds suggest? What could go wrong in the experiment to make the bounds too optimistic

Answer:

There were some assumptions and simplifications made in 1.1 that may require more test samples to be generated then are necessary. This could occur if some unexpected factor influences the click rate of one of the colors, making one a clear winner over the other choices from the beginning. One reason for this could be violation of the independence assumption for the distributions of click rates given a specific color. A few reasons why my tail bound estimates might be too optimistic (tail is fatter than expected): -The variance is not bounded by a hard upper limit at 7%, as previously assumed. -Uneven distribution of web activity (spikes of high click rates outweigh low click rates, such as how Google activity is likely maximized during daylight hours in the US. There would have to be a time-based weight assigned to click instances to account for this.

1.3 Radford Neal's Priors, [Bonus 20 pts]

Assume that we have some function

$$f_n(x) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \alpha_i \phi_i(x)$$

that is being evaluated on points $x_1, ...x_m$. Assume that the weights α_i are drawn iid from the uniform distribution $U[-1, 1]$, and that all ϕ_i are bounded in their L_∞ norm, i.e. $|\phi_i(x)| \leq C$ for some constant C.

Analyze the convergence of the vector $(f_n(x_1)...f_n(x_m))$ as $n \to \infty$. Hint: you can invoke the central limit theorem. Can you derive a condition on the covariance between two function values $f_n(x_i)f_n(x_j)$?

Answer:

We can apply the central limit theorem,

$$f_n(x) \sim N(0, \frac{1}{3} \sum_i \phi_i^2(x)) \tag{4}$$

where $\frac{1}{3}$ is the variance of the uniform distribution of α_i on the range [-1,1]. The covariance between $f(x_i)$ and $f(x_j)$ can be derived as follows (assuming that the mean of f(x) is 0):

$$
\begin{aligned}
E[f(x_i)f(x_j)] &= \frac{1}{n}E[\sum_k \alpha_{ik}\phi_{ik}(x_i)\sum_i \alpha_{ji}\phi_{ji}(x_j)] \\
&= \frac{1}{n}\sum_k\sum_i E[\alpha_{ik}\alpha ji\phi_{ji}(x_j)] \\
&= \frac{1}{n}\sum_k\sum_l E[\alpha_{ik}^2\phi_{ik}(x_i)^2] \\
&= \frac{1}{n}\sum_k \phi_{ik}(x_i)^2 E[\alpha_k^2] \\
&= \frac{1}{3n}\sum_k \phi_{ik}(x_i)^2 \\
&\leq \frac{C^2}{3}
\end{aligned}
$$

2 K Nearest Neighbours (KNN) Regression [Di; 50pts]

Let's define the following notations for a KNN regression problem. Let **k** be the number of neighbours to be considered, and **n** be the number of entries in the training data. Denote by **X** the input space and by **Y** the scalar output space. Then the i.i.d. training data can be listed as $D = (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) \in \mathbf{X} \times \mathbf{Y}$. There is an underlying joint probability distribution μ on $\mathbf{X} \times \mathbf{Y}$, but we never know it in practice. Instead we only try to learn a function $\hat{f} : \mathbf{X} \to \mathbf{Y}$ from the data D to approximate conditional behavior of μ given X.

2.1 KNN v.s. Boxcar Kernel [10+10 pts]

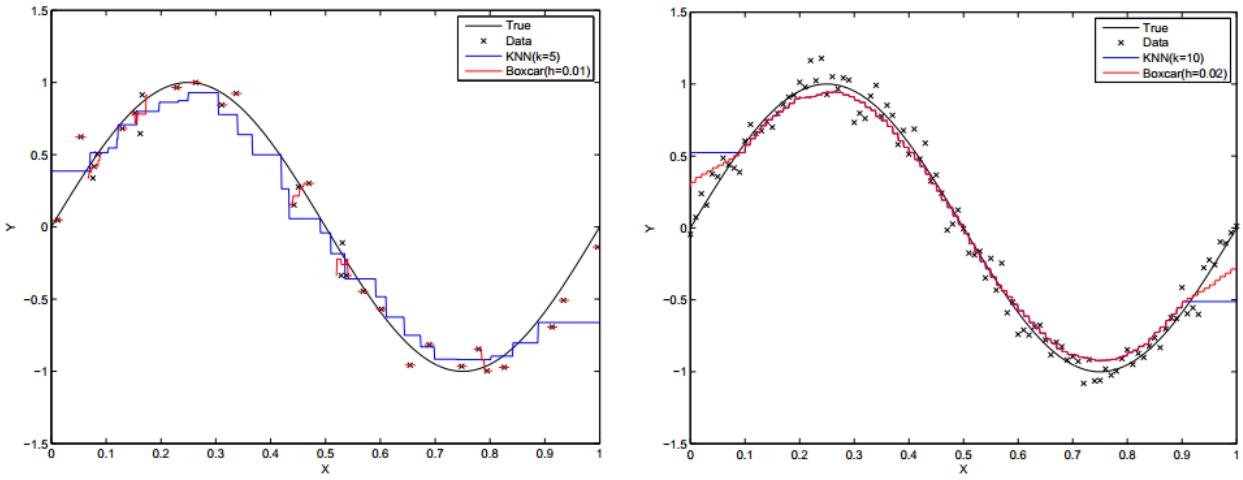
KNN regression is similar to KNN classification. Intuitively, after we identify the k nearest neighbours, we draw an estimation from the the weighted average of the neighbours. On the other hand, when using Boxcar Kernel to do Kernel Regression, we instead consider the weighted average of all the neighbours confined within a certain bandwidth (**h**) centered around the target.

1. Draw 2 plots and argue why in certain conditions KNN provides better estimations than Boxcar kernel and vice-versa.

   SOLUTION: Several students in our class made very nice plots.

Figure 1 shows regression results based on KNN and Boxcar Kernel for two cases. The true curve is given by $Y = \sin(2\pi X)$. In the case shown in Figure 1 (a), we have 30 data points which are unevenly distributed in the X domain. In the figure, KNN regression curve with $k = 5$ and Boxcar kernel regression with $h = 0.01$ are plotted. As seen in the figure, KNN tracks the trend of the true curve while the boxcar kernel regression can not make a continuous curve because of the sparse data and the small window size.

In Figure 1 (b), we have evenly distributed 100 data point. Both KNN ($k = 10$) and Boxcar kernel ($h = 0.02$) show good estimate of the underlying curve. The accuracy of the estimates around the $0.15 < X < 0.85$ are almost the same; however, Boxcar kernel shows better estimation accuracy around the boundaries than KNN. This is because Boxcar kernel takes weighted average of all data within its window while the regression result of KNN is more influenced by the inner side data.



(a) Sparse data and small window size (b) Dense data and sufficient window size

Figure 1: Comparisons of KNN regression and Boxcar kernel regression. The true curve is given by $Y = \sin(2\pi X)$.

Figure 1: Version 1

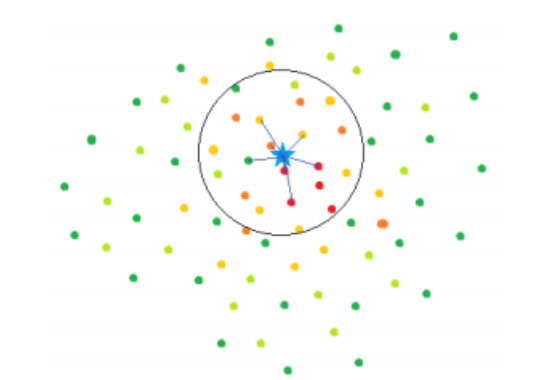


Figure 1: A case where the boxcar kernel performs better.

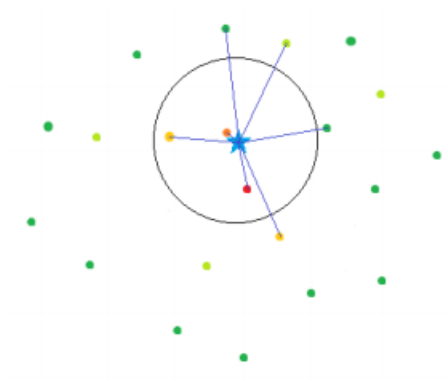


Figure 2: A case where the KNN performs better.

Figure 2: Version 2

2.2 Bias Variance Trade-off [5+15+10 pts]

Commonly we use the squared loss to measure the expected prediction error of $\hat{f}(X)$ over the labelled training data, i.e. $\mathbb{E}\left[(Y-\hat{f}(x))^2|X=x\right]$. This quantity is minimized at $f(x)=\mathbb{E}\left[Y|X=x\right]$. This $f(x)$ is called the true regression function for **X** and **Y**, which is also unknown in practice. Notice that although $f(x)$ minimize the prediction error, but it does NOT reduce it to zero. Therefore, we write $Y=f(X)+\epsilon$, where ϵ has mean of zero and variance of σ^2. Notice that σ^2 is called the Bayes error.

1. Where does the Bayes error come from? Is it preventable?

   SOLUTION: It come from noise in data, which is not preventable. In fact Bayes error can also be referred to as Irreducible Error.

2. Expand the formula of the Expected Prediction Error, i.e., $\mathbb{E}\left[(Y-\hat{f}(x))^2|X=x\right]$ in terms of the Bayes error and the mean squared error of $\hat{f}(x)$ in estimating $f(x)$. Then, expand further to show how the latter consists of **bias** and **variance** of $\hat{f}$ w.r.t. f.

   SOLUTION: One student made an excellent point by pointing something out, which is very subtle and often neglected. I hereby honor his solution.

$$E\left[\left(Y-\hat{f}(x)\right)^2\right]=E\left[\left(f(X)+\epsilon-\hat{f}(x)\right)^2\right]$$

$$=E\left[\left(f(X)-\hat{f}(x)\right)^2\right]+2E\left[\epsilon\left(f(x)-\hat{f}(x)\right)\right]+E[\epsilon^2]=MSE\left(f,\hat{f}\right)+Var(\epsilon)$$

Assuming the Bayes error is independent of $\hat{f}(x)$,

$$E\left[\epsilon\left(f(x)-\hat{f}(x)\right)\right]=E[\epsilon]E\left[f(x)-\hat{f}(x)\right]=0$$

$$E[\epsilon^2]=\sigma^2+E[\epsilon]^2=\sigma^2$$

$$E\left[\left(f(X)-\hat{f}(x)\right)^2\right]=E\left[\left(\left(f(X)-E[\hat{f}(x)]\right)+\left(E[\hat{f}(x)]-\hat{f}(x)\right)\right)^2\right]$$

$$=E\left[\left(f(X)-E[\hat{f}(x)]\right)^2+2\left(f(X)-E[\hat{f}(x)]\right)\left(E[\hat{f}(x)]-\hat{f}(x)\right)+\left(E[\hat{f}(x)]-\hat{f}(x)\right)^2\right]$$

$$=E\left[\left(f(X)-E[\hat{f}(x)]\right)^2\right]+2E\left[\left(f(X)-E[\hat{f}(x)]\right)\left(E[\hat{f}(x)]-\hat{f}(x)\right)\right]+E\left[\left(E[\hat{f}(x)]-\hat{f}(x)\right)^2\right]$$

We can show:

$$2E\left[\left(f(X)-E[\hat{f}(x)]\right)\left(E[\hat{f}(x)]-\hat{f}(x)\right)\right]=2\left(f(X)-E[\hat{f}(x)]\right)E\left[E[\hat{f}(x)]-\hat{f}(x)\right]=0$$

Finally,

$$E\left[\left(f(X)-\hat{f}(x)\right)^2\right]=E\left[\left(f(X)-E[\hat{f}(x)]\right)^2\right]+E\left[\left(E[\hat{f}(x)]-\hat{f}(x)\right)^2\right]$$

$$=Bias(f(x),\hat{f}(x))^2+Var(\hat{f}(x))$$

Putting it all together:

$$E\left[\left(Y-\hat{f}(x)\right)^2\right]=Bias(f(x),\hat{f}(x))^2+Var\left(\hat{f}(x)\right)+\sigma^2$$

3. Describe how the bias and variance change w.r.t. **k** in KNN Regression, and the bandwidth **h** in Kernel Regression.

   SOLUTION: As k or h increases, the variance goes down, but the bias increases. In extreme case, consider k=n, and h becomes infinity.

2.3 Consistency [Bonus 5+10+5 pts ... It's EASY!!!]

1. What is consistency?

   SOLUTION: An estimator is consistent if given infinitely many samples, it converges to the truth in probability. Formally. $P(|\hat{\theta_n} - \theta| > \sigma) \to 0, n \to \infty$.

2. Under what conditions is KNN regression consistent? Argue in terms of **k** and **n**. Please also give a mathematical relation between **k** and **n**.

   SOLUTION: $k \to \infty, n \to \infty, \frac{k}{n} \to 0$. For example, $k = log(n)$

3. Under what condition is the consistency of KNN regression not guaranteed, regardless of **k** or **n**?

   SOLUTION: It will not hold in infinite-dimensional spaces.