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?

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?

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 $f_n(x_i)f_n(x_j)$?

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

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.

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 $f(x) = \mathbb{E}[Y|X = x]$. 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?
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.
3. Describe how the bias and variance change w.r.t. **k** in KNN Regression, and the bandwidth **h** in Kernel Regression.

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

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