Homework 4

November 2, 2016

* For each question, give your answer. Justify your answer using a detailed explanation.

1 Bias-variance decomposition

For a random variable z, let $\bar{z}$ denote its mean. Suppose our observation is generated by the true function f as

$$y = f(x) + \epsilon$$

where ϵ is normally distributed with zero mean and standard deviation σ. Training data set is $\mathcal{D} = \{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i=1}^{m}$, from which you learned your hypothesis function $h_{\mathcal{D}}$.The bias-variance tradeoff is an important aspect of data science projects based on machine learning.

Now for a new data point x^* out of $\mathcal{D}$, we want to investigate the expected error between the predicted value and the observation y^*, i.e.,

$$\mathbb{E}_{\mathcal{D},\epsilon}[\ (y^* - h(x^*))^2\ ].$$

This error can be decomposed into three parts namely: `variance, bias`^{2}, and `noise`, where the expectation is taken over all possible training set $\mathcal{D}$. Here

$$\begin{aligned} \texttt{variance} &= \mathbb{E}_{\mathcal{D}}[\ \left(h(x^*) - \overline{h(x^*)}\right)^2\ ] \\ \texttt{bias}^2 &= [\ \overline{h(x^*)} - f(x^*)\ ]^2 \\ \texttt{noise} &= \sigma^2 \end{aligned}$$

and $\overline{h(x^*)} = \mathbb{E}_{\mathcal{D}} h(x^*)$.

The error `bias` is the amount by which the expected model prediction differs from the true value or target; while `variance` measures how inconsistent are the predictions from

one another, over different training sets, not whether they are accurate or not. **Models that exhibit small variance and high bias underfit the truth target. Models that exhibit high variance and low bias overfit the truth target.**

The data scientist's goal is to simultaneously reduce bias and variance as much as possible in order to obtain as accurate model as is feasible. However, there is a tradeoff to be made when selecting models of different flexibility or complexity and in selecting appropriate training sets to minimize these sources of error.

Question:

Show that

$$\mathbb{E}_{\mathcal{D},\epsilon}[\ (y^* - h(x^*))^2\ ] = \texttt{variance} + \texttt{bias}^2 + \sigma^2.$$

Hints: first you may want to prove a lemma that for any random variable, it holds true that

$$\mathbb{E}[(z - \bar{z})^2] = \mathbb{E}[z^2] - \bar{z}^2,$$

so that

$$\mathbb{E}[z^2] = \mathbb{E}[(z - \bar{z})^2 + \bar{z}^2.$$

It follows that $(y^* - h(x^*))^2 = (y^*)^2 - 2h(x^*)y^* + h(x^*)^2$. Note that y^* and $h(x^*)$ are independent variables. The result follows from using the lemma twice. Also note $\mathbb{E}_{\epsilon}[\ (y^* - f(x^*))^2\ ] = \sigma^2$.

2 Ridge regression

Ridge regression has two versions. One is the regularized version:

$$\min_{\theta} \quad \frac{1}{2}\|\mathbf{y} - \Phi\theta\|_2^2 + \frac{\mu}{2}\|\theta\|_2^2, \tag{2.1}$$

and the other is constrained version

$$\begin{array}{ll} \min\limits_{\theta} & \frac{1}{2}\|\mathbf{y} - \Phi\theta\|_2^2 \\ \text{s.t.} & \|\theta\|_2^2 \le C, \end{array} \tag{2.2}$$

with $C \ge 0$ and $\lambda \ge 0$ are given parameters.

Questions:

1. For any given $\mu \in [0, +\infty)$, find the optimal solution θ^{R1} to (2.1).

2. Find the optimal solution θ^{R2} to (2.2) for any given $C \in [0, +\infty)$. (We've done the part for $C \geq \|\theta^{LS}\|_2^2$ in class using KKT conditions, where $\theta^{LS} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$ be the LS solution.)

3. For given $\mu \in [0, +\infty)$, you have the optimal solution θ^{R1} to (2.1). Now explain for what value of C (you may want to use θ^{R1} to determine C), the optimal solution θ^{R2} to (2.2) is equivalent to θ^{R1}, i.e., $\theta^{R2} = \theta^{R1}$.

4. For given $C \in (0, +\infty)$, you have the optimal solution θ^{R2} to (2.2). Now explain for what value of μ (you may want to use θ^{R2} to determine λ), (2.1) yields the same solution as (2.2), i.e., $\theta^{R1} = \theta^{R2}$.

3 Nonlinear Transformation, overfitting, and validation

A consumer price index (CPI) measures changes in the price level of a market basket of consumer goods and services purchased by households. The annual percentage change in a CPI is used as a measure of inflation. A CPI can be used to index (i.e., adjust for the effect of inflation) the real value of wages, salaries, pensions, for regulating prices and for deflating monetary magnitudes to show changes in real values. Generally, a CPI $\geq 3\%$ implies inflation, and a CPI $\geq 5\%$ indicates serious inflation. For a single item, the CPI is calculated by

$$\frac{CPI_2}{CPI_1} = \frac{Price_2}{Price_2}$$

where 1 means the comparison time period and CPI_1 is usually considered as an index of 100%. For multiple items, the overall CPI is given by

$$CPI = \frac{\sum_{i=1}^{n} CPI_i \times weight_i}{\sum_{i=1}^{n} weight_i}$$

where the $weight_i$s do not necessarily sum up to 1 or 100.

You are given the following data sets Table 1 and 2 about the monthly CPI of China[1].

[1] http://www.inflation.eu

Table 1: Monthly CPI of China, 2015. Increase as prior month （较上月增长）

Time	CPI
January 2015 - December 2014	0.26%
February 2015 - January 2015	1.23%
March 2015 - February 2015	−0.52%
April 2015 - March 2015	−0.26%
May 2015 - April 2015	−0.17%
June 2015 - May 2015	0.00%
July 2015 - June 2015	0.35%
August 2015 - July 2015	0.52%
September 2015 - August 2015	0.09%
October 2015 - September 2015	−0.35%
November 2015 - October 2015	0.00%
December 2015 - November 2015	0.52%

Table 2: Monthly CPI of China, 2015. Increase as prior month （较同期增长）

Time	CPI
January 2015 - January 2014	0.74%
February 2015 - February 2014	1.41%
March 2015 - March 2014	1.45%
April 2015 - April 2014	1.49%
May 2015 - May 2014	1.21%
June 2015 - June 2014	1.31%
July 2015 - July 2014	1.68%
August 2015 - August 2014	2.04%
September 2015 - September 2014	1.58%
October 2015 - October 2014	1.23%
November 2015 - November 2014	1.50%
December 2015 - December 2014	1.67%

Questions: (you can choose **either one** of the two tables above to do following homework):

1. For polynomial regression model of order $n = 0, 1, 2, ..., 9$, find the optimal parameter values for **each** model, and complete the following table (for example, if you choose quadratic model, it would be written as

$$CPI = \theta_0 + \theta_1 \times Time + \theta_2 \times Time^2$$

where θ_0 and θ_1 are your parameters). For each model, you can directly solve the normal equation by Matlab linear equation solver. (`inv` is not recommended)

n	0	1	2	3	4	5	6	7	8	9
θ_0										
θ_1										
θ_2										
θ_3										
θ_4										
θ_5										
θ_6										
θ_7										
θ_8										
θ_9										

2. Write down the condition number for each $\Phi^T\Phi$, $n = 0, 1, ..., 9$.

3. Plot the change of $\|\theta\|^2$ and the residual ($E_{in}(\theta)$ for each model, $n = 0, ..., 9$).

4. Consider the polynomial model of order 10 and the corresponding ridge regression

$$\min_{\theta} \quad \frac{1}{2}\|\Phi\theta - y\|_2^2 + \frac{\lambda}{2}\|\theta\|_2^2.$$

   For $\lambda = 0, 0.01, 0.02, ..., 1$, use LOOCV to calculate an estimate of E_{out}. Plot E_{out}, and find the best regularization parameter value you found.