CS 189 Introduction to Machine Learning
Fall 2017 DIS3

1 Projection, Approximation, and Estimation

In this discussion, we will revisit a fundamental issue that ought to have bothered you throughout the class so far. In typical applications, we are dealing with data generated by an *unknown* function (with some noise), and our goal is to estimate this unknown function from samples. So far, we have used linear and polynomial regression as our only methods, but are these good methods when the function is not a polynomial?

We will answer a few aspects of this general question. In particular, we will provide geometric answers to:

- Can we do least squares on an arbitrary problem? What do we end up doing in this case?
- How large does the degree have to be for us to execute reliable polynomial regression?
- Can we formulate the bias-variance trade-off of polynomial regression?

Doing this discussion in sequence will set up all the necessary tools you need to analyze the problem. You are recommended to go through this discussion using the parts (a, e, f, g) first, recalling the geometric intuition conveyed during discussion. Draw pictures of what these projections are doing; that's a great way to develop intuition for what is going on!

———-

Define the projection of a vector $y \in \mathbb{R}^n$ onto a (closed) set $\mathscr{C}$ as

$$P_{\mathscr{C}}(y) = \arg\min_{u \in \mathscr{C}} \|y-u\|_2^2.$$

For a matrix $A \in \mathbb{R}^{n \times d}$ having full column rank, let the set $c(A)$ denote its column space.

(a) Derive a closed form expression for $P_{c(A)}(y)$ in terms of the matrix A and vector y.

(b) From this problem onwards, we use the shorthand $P_{c(A)}(y) = P_A(y)$. For arbitrary vectors $y^* \in \mathbb{R}^n$ and $w \in \mathbb{R}^n$, show that

$$\|y^* - P_A(y^*+w)\|_2 \leq \|y^* - P_A(y^*)\|_2 + \|P_A(w)\|_2.$$

Hint: The triangle inequality $\|a+b\|_2 \leq \|a\|_2 + \|b\|_2$ may be useful.

(c) Furthermore, show that

$$\|y^* - P_A(y^*+w)\|_2^2 = \|y^* - P_A(y^*)\|_2^2 + \|P_A(w)\|_2^2.$$

Hint: Recall the Pythagorean theorem.

(d) Let us introduce the shorthand $y_P^* = P_A(y^*)$. Use the previous part to show that

$$\|y^* - P_A(y^* + w)\|_2^2 = \|y^* - y_P^*\|_2^2 + \|y_P^* - P_A(y_P^* + w)\|_2^2.$$

Hint: What is the projection $P_{\mathscr{C}}(y)$ when $y \in \mathscr{C}$?

(e) Let $x^* = \arg\min_x \|y^* - Ax\|_2^2$, and let $\hat{x} = \arg\min_x \|Ax^* + w - Ax\|_2^2$. Note that $\hat{x}$ is a random variable (since it is a function of the random noise w), and that $Ax^* = y_P^*$. Show that

$$\mathbb{E}\left[\|y^* - P_A(y^* + w)\|_2^2\right] = \underbrace{\|y^* - Ax^*\|_2^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[\|Ax^* - A\hat{x}\|_2^2\right]}_{\text{variance}}.$$

Conclude that the error of estimating an arbitrary vector y^* corrupted by noise via linear regression is bounded by the sum of two terms i) an approximation error, which captures how far y^* is from the assumed linear model, and ii) an estimation error term, which captures the error made if the model were indeed linear.

(f) You will see in HW3 that $\frac{1}{n}\mathbb{E}\left[\|Ax^* - A\hat{x}\|_2^2\right] = d/n$ when A is a full-rank $n \times d$ matrix and $w \sim \mathscr{N}(0,1)$.

Let us say that we obtain n samples $\{x_i, y_i\}_{i=1}^n$, where $y_i = e^{x_i} + w_i$. Here, each point $x_i \in [-3,3]$ is distinct, and each $w_i \overset{i.i.d.}{\sim} \mathscr{N}(0,1)$ represents independent random noise. Since the function is unknown to us a-priori, we decide to use polynomial regression with degree D to estimate the relationship between x_i and y_i. Stack up the noiseless function evaluations into the vector y^*, whose ith coordinate is given by $y_i^* = e^{x_i}$.

Using Taylor expansion (HW2 may be useful) and the above parts, show that if our estimate $\hat{y}$ is obtained by performing least squares, then we have

$$\frac{1}{n}\mathbb{E}\left[\|y^* - \hat{y}\|_2\right] \leq e^3 \frac{3^{D+1}}{(D+1)!} + \frac{D+1}{n}.$$

(g) In part (f), notice that as D increases, the approximation error decreases but the estimation error increases. Discuss qualitatively why that is true. Given n samples, show that setting $D = O(\log n / \log\log n)$ is an optimal choice for this problem.