CS 189 Fall 2017 — Introduction to Machine Learning — 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.

Solution: By definition, the projection $P_{c(A)}(y)$ is given by the solution to the least squares problem $P_{c(A)}(y) = A \arg\min_x \|v - Ax\|_2^2$. Using the solution to the least squares problem, we can simplify this as

$$P_{c(A)}(y) = A(A^\top A)^{-1} A^\top y.$$

(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.

Solution: Notice that for two vectors u and v, the projection operation is given by

$$P_A(u+v) = A(A^\top A)^{-1}A^\top(u+v) = A(A^\top A)^{-1}A^\top u + A(A^\top A)^{-1}A^\top v = P_A(u) + P_A(v),$$

and so the projection operation onto the column space of A is linear.

Consequently, we have

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

where the final step uses the triangle inequality.

(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.

Solution: Recall that the error vector $y^* - P_A(y^*)$ is orthogonal to any vector in the column space of A (revise the geometry of least squares if this is not clear to you).

Hence, by the Pythagorean theorem, we have

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

(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 $y \in \mathscr{C}$?

Solution: Notice that by definition, $P_{\mathscr{C}}(y) = y$ when $y \in \mathscr{C}$. Hence, by the linearity of projections, we may write

$$\begin{aligned} P_A(y_P^* + w) &= P_A(y_P^*) + P_A(w) \\ &= P_A(y^*) + P_A(w). \end{aligned}$$

Using part (c) proves the required result.

(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.

Solution: By the definition of the projection, we have $\|y^* - y_P^*\|_2 = \min_{y \in c(A)} \|y^* - y\|_2 = \min_x \|y^* - Ax\|_2 = \|y^* - Ax^*\|_2$.

Similarly, $\|y_P^* - P_A(y_P^* + w)\|_2 = \|Ax^* - P_A(Ax^* + w)\|_2 = \|Ax^* - A\hat{x}\|_2$.

Substituting these terms into the solution of part (d) proves the result.

The first term in this expansion is the approximation error term, capturing how far y^* is from the column space of A. The second term is the estimation term, showing how well we can estimate a linear model.

(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 $w \sim \mathcal{N}(0,1)$.

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 $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ 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

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

Solution: Notice that here, we are performing polynomial regression with a matrix A of column rank $D+1$, since all the points are distinct (recall HW2 problem 3). We must evaluate the approximation error and estimation error defined in the previous part. From HW3, we know that the estimation error is given by

$$\frac{1}{n}\mathbb{E}\left[\|Ax^* - A\hat{x}\|_2^2\right] = \frac{D+1}{n}.$$

For the approximation error, we use the Dth order Taylor series approximation of e^x about the point 0, which is given (for some $x' \in [0,x]$) by

$$e^x = \underbrace{\sum_{i=0}^{D} \frac{x^i}{i!}}_{\phi_D(x)} + e^{x'} \frac{x^{D+1}}{(D+1)!}.$$

Now notice that for all $x \in [-3,3]$, we have the relation $|e^x - \phi_D(x)| \leq e^3 \frac{3^{D+1}}{(D+1)!}$. Consequently, the approximation error for the sample $\{x_i, y_i^*\}$ is bounded by $e^3 \frac{3^{D+1}}{(D+1)!}$. Summing and dividing through by n yields the result.

(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.

Solution: The approximation error measures how well our regression model fits the true data. Since we can better approximate the unknown function with higher order polynomials, this term decreases with D. However, it becomes harder to estimate the correct polynomial, since we now give ourselves extra freedom. This is the bias variance tradeoff in action!

Equating the two terms in the upper bound of part (f), we have

$$e^3 \frac{3^{D+1}}{(D+1)!} = \frac{(D+1)}{n}.$$

You have seen such an equation before (in HW2, problem 2). The choice of D hinges on the same ideas (Stirling's approximation, and guessing a solution that is within a constant factor of being optimal). This leads us to the choice $D = O(\log n / \log \log n)$.