Homework 6

36-462/662, Data Mining, Fall 2019

Due at 10 pm on Wednesday, 9 October 2019

AGENDA: Hammering home the importance of not evaluating predictive models on testing data.

1. *Online questions* (10) are online and due on Sunday at 10 pm.

2. *Optimism* If we use data $(X_1, Y_1), \ldots (X_n, Y_n)$ to learn a predictive model $\hat{\mu}$, the "optimism" of our method is defined as how much worse that model would do no new data with the same values of X but *independent* Ys. That is, for each i, Y_i' has the same distribution as Y_i (conditional on X_i), but is independent of Y_i, and the optimism (for regression) is

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i' - \hat{\mu}(X_i))^2\right] - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}(X_i))^2\right] \quad (1)$$

In this problem, we'll see how to build a simple, unbiased estimator of the optimism.

(a) (5) Show that the optimism (as defined above) is equal to

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(Y_i' - \hat{\mu}(X_i))^2\right] - \mathbb{E}\left[(Y_i - \hat{\mu}(X_i))^2\right] \quad (2)$$

(b) (5) Show that $\mathbb{E}\left[Y_i' - \hat{\mu}(X_i)\right] = \mathbb{E}\left[Y_i - \hat{\mu}(X_i)\right]$.

(c) (5) Show that the optimism is equal to

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}\left[Y_i' - \hat{\mu}(X_i)\right] - \mathrm{Var}\left[Y_i - \hat{\mu}(X_i)\right] \quad (3)$$

(d) (5) Show that

$$\mathrm{Var}\left[Y_i - \hat{\mu}(X_i)\right] = \mathrm{Var}\left[Y_i\right] + \mathrm{Var}\left[\hat{\mu}(X_i)\right] - 2\mathrm{Cov}\left[Y_i, \hat{\mu}(X_i)\right] \quad (4)$$

(e) (5) Show that

$$\mathrm{Var}\left[Y_i' - \hat{\mu}(X_i)\right] = \mathrm{Var}\left[Y_i\right] + \mathrm{Var}\left[\hat{\mu}(X_i)\right] \quad (5)$$

(f) (5) Show that the optimism equals

$$\frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}\left[Y_i, \hat{\mu}(X_i)\right] \tag{6}$$

(g) (5) From this point, assume that our predictor is a linear smoother, so that $\hat{\mu}(X_i) = \sum_{j=1}^{n} w(X_i, X_j)Y_j$, or just $\sum_{j=1}^{n} w_{ij}Y_j$ for short. Find an expression for $\mathrm{Cov}\left[Y_i, \hat{\mu}(X_i)\right]$ in terms of the matrix $\mathbf{w}$ and the variance matrix of the Y's.

(h) (5) From this point, assume that $Y_i = \mu(X_i) + \epsilon_i$, with $\mathbb{E}\left[\epsilon_i | X_i\right] = 0$, $\mathrm{Var}\left[\epsilon_i | X_i\right] = \sigma^2$, and $\mathrm{Cov}\left[\epsilon_i, \epsilon_j\right] = 0$ when $i \neq j$. Show that $\mathrm{Cov}\left[Y_i, \hat{\mu}(X_i)\right] = \sigma^2 w_{ii}$. (It's possible to do this part without doing the previous one, but if you did the previous one right, your general formula should reduce to this under the extra assumptions.)

(i) (5) Show that, under all these assumptions, the optimism is

$$\frac{2\sigma^2}{n}\,\mathrm{tr}\,\mathbf{w} \tag{7}$$

(j) (5) What is the optimism of a linear regression model with p coefficients? Answer in terms of σ^2, n and p (and numerical constants such as 2 or π). (*Hint:* Homework 2.) Does the optimism $\to 0$ as $n \to \infty$ with p fixed??

(k) (5) What is the optimism of a k-nearest-neighbor regression? Answer in terms of σ^2, n and k (and numerical constants). Does the optimism $\to 0$ as $n \to \infty$ with k fixed?

3. *Leave-one-out* In class, we saw a short-cut formula for leave-one-out cross-validation for linear smoothers, where $\hat{\mu}(X) = \sum_{j=1}^{n} w(x, x_j)y_j$, namely

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}(x_i)}{1 - w_{ii}}\right)^2 \tag{8}$$

(a) (5) Explain why $w_{ii} = 1/k$ for k-nearest-neighbor regression.

(b) (5) Explain why the short-cut formula simplifies to multiplying the in-sample error by $\left(\frac{k}{k-1}\right)^2$.

(Is there something funny about all this when $k = 1$?)

4. (a) (5) Explain what the following code does.

```
sim.poly <- function(n, degree) {
    x <- runif(n, min=-2, max=2)
    poly.x <- poly(x, degree=degree, raw=TRUE)
    alternating.signs <- rep(c(-1,1),length.out=degree)
```

```
    sum.poly <- poly.x %*% alternating.signs
    y <- sum.poly+rnorm(n,0,0.1)
    return(data.frame(x=x,y=y))
}
```

(You might need to look up the `poly()` function.)

(b) (2) Use the code from the previous part to generate 3 data frames, where Y is a quadratic function of X plus noise. The data frames should contain 100, 1000 and 10000 rows. Check that they have the right dimensions, and that each one shows the expected quadratic relationship between X and Y.

(c) (8) For each of the three data sets, do k-nearest-neighbor regression with k running from 1 to 100. Plot both the in-sample error and the generalization error as estimated by leave-one-out cross-validation. What k is the best (by LOOCV) at each sample size? Why does it change with n? Plot the predictions you get from the selected k — are they getting visibly better?

Rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is properly integrated with a tool like R Markdown or knitr, and both the knitted file and the source file are submitted. The code is indented, commented, and uses meaningful names. All code is relevant; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and raw computer code or output are only shown when explicitly asked for.

Extra credit (5): Both LOOCV and the optimism formula give ways of estimating the generalization error. When are they compatible, that is, when do they give (approximately) the same estimate of the generalization error?