Stats216v: Statistical Learning

Stanford University
Summer 2017
Gyu-Ho Lee (gyuhox@gmail.com (mailto:gyuhox@gmail.com))

1. Introduction

The Supervised Learning Problem

- Outcome measurement Y (also called dependent variable, response, target).
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
- In the regression problem, Y is quantitative (e.g price, blood pressure).
- In the classification problem, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
- We have training data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.

On the basis of the training data we would like to:

- Accurately predict unseen test cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences.

Unsupervised learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy
  - find groups of samples that behave similarly
  - find features that behave similarly
  - find linear combinations of features with the most variation.
- difficult to know how well your are doing.
- different from supervised learning, but can be useful as a pre-processing step for supervised learning.

Statistical Learning versus Machine Learning

- Machine learning arose as a subfield of Artificial Intelligence.
- Statistical learning arose as a subfield of Statistics.
- There is much overlap — both fields focus on supervised and unsupervised problems:
- Machine learning has a greater emphasis on large scale applications and prediction accuracy.
- Statistical learning emphasizes models and their interpretability, and precision and uncertainty.
- But the distinction has become more and more blurred, and there is a great deal of “cross-fertilization”.
- Machine learning has the upper hand in Marketing!

1.2.R1

Which of the following are supervised learning problems? More than one box can be checked.

1. Predict whether a website user will click on an ad.
2. Find clusters of genes that interact with each other.
3. Classify a handwritten digit as 0-9 from labeled examples.
4. Find stocks that are likely to rise.

Gyu-Ho's Answer: 1, 3, 4.

Problems with clearly defined "predictors" and "responses" are supervised.

1.2.R2

True or False: The only goal of any supervised learning study is to be able to predict the response very accurately.

Gyu-Ho's Answer: False. Not just for "prediction", also for "inference".

False. Most supervised learning problems can be framed formally in terms of predicting a response, but prediction alone is often not the main goal of the analysis. For example, many applications of linear regression in the sciences are aimed primarily at understanding how the inputs of a system drive the outputs; an extremely complicated "black box" giving pure predictions would not be very useful in and of itself.

2. Statistical Learning

2.1 Introduction to Regression Models

Now we write our model as $Y = f(X) + \varepsilon$. With a good f we can make predictions of Y at new points $X = x$.

$f(4) = E(Y|X = 4)$ means expected value (average) of Y given $X = 4$. This ideal $f(x) = E(Y|X = x)$ is called the regression function.

The ideal or optimal predictor of Y with regard to mean-squared prediction error: $f(x) = E(Y|X = x)$ is the function that minimizes $E[(Y - g(X))^2|X = x]$ over all functions g at all points $X = x$.

$\varepsilon = Y - f(x)$ is the **irreducible** error — i.e. even if we knew $f(x)$, we would still make errors in prediction, since at each $X = x$ there is typically a distribution of possible Y values.

For any estimate $\hat{f}(x)$ of $f(x)$, we have

$$E[(Y - \hat{f}(X))^2|X = x] = [f(x) - \hat{f}(x)]^2 + Var(\varepsilon)$$

.

Typically we have few if any data points with $X = 4$ exactly. So we cannot compute $E(Y|X = x)$.

Relax the definition and let

$$f(x) = Ave(Y|X \in N(x))$$

where $N(x)$ is some neighborhood of x.

Nearest neighbor methods can be lousy when p is large. Reason: the curse of dimensionality. Nearest neighbors tend to be far away in high dimensions.

2.1.R1

In the expression $Sales \approx f(TV, Radio, Newspaper)$, $Sales$ is the:

1. Response
2. Training Data
3. Independent Variable
4. Feature

Gyu-Ho's Answer: Response.

The variable which you are trying to model is called the response or outcome. The other variables are called features, predictors, or independent variables. Together, the collection of features and response values that you will use for fitting form your training data.

2.2.R1

A hypercube with side length 1 in d dimensions is defined to be the set of points $(x_1, x_2, \ldots, x_d)$ such that $0 \leq x_j \leq 1$ for all $j = 1, 2, \ldots, d$. The boundary of the hypercube is defined to be the set of all points such that there exists a j for which $0 \leq x_j \leq .05$ or $.95 \leq x_j \leq 1$ (namely, the boundary is the set of all points that have at least one dimension in the most extreme 10 of possible values). What proportion of the points in a hypercube of dimension 50 are in the boundary? (hint: you may want to calculate the volume of the non-boundary region)

Please give your answer as a value between 0 and 1 with 3 significant digits. If you think the answer is 50.52, you should say 0.505:

Gyu-Ho's Answer: 0.995

- The volume of hypercube with 50-dimension and side length 1 is $1^{50} = 1$.
- The volume of hypercube interior is $0.9^{50} = 0.005$.
- Thus, the volume of boundary is $1 - 0.005 = 0.995$.

2.3.R1

True or False: A fitted model with more predictors will necessarily have a lower Training Set Error than a model with fewer predictors.

Gyu-Ho's Answer: False.

False. While we typically expect a model with more predictors to have lower Training Set Error, it is not necessarily the case. An extreme counterexample would be a case where you have a model with one predictor that is always equal to the response, compared to a model with many predictors that are random.

2.3.R2

While doing a homework assignment, you fit a Linear Model to your data set. You are thinking about changing the Linear Model to a Quadratic one. Which of the following is most likely true:

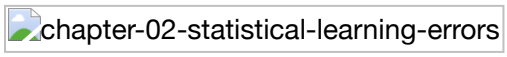
1. Using the Quadratic Model will decrease your Irreducible Error.
2. Using the Quadratic Model will decrease the Bias of your model.
3. Using the Quadratic Model will decrease the Variance of your model.
4. Using the Quadratic Model will decrease your Reducible Error.

Gyu-Ho's Answer: Using the Quadratic Model will decrease the Bias of your model, because it's more flexible.

Introducing the quadratic term will make your model more complicated. More complicated models typically have lower bias at the cost of higher variance. This has an unclear effect on Reducible Error (could go up or down) and no effect on Irreducible Error.

2.4.R1

Look at the graph given on page 30 of the Chapter 2 lecture slides. Which of the following is most likely true of what would happen to the Test Error curve as we move $1/K$ further above 1?



1. The Test Errors will increase.
2. The Test Errors will decrease.
3. Not enough information is given to decide.
4. It does not make sense to have $1/K > 1$.

Gyu-Ho's Answer: It does not make sense to have $1/K > 1$.

Since K is the number of neighbors, the value of K must be a Natural Number. This means that $1/K \leq 1$.

2.Q

For each of the following parts, indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible model.

- Flexible is better.
- Flexible is worse.

2.Q.1

The sample size n is extremely large, and the number of predictors p is small:

Gyu-Ho's Answer: Flexible is better, when we have much data.

A flexible model will allow us to take full advantage of our large sample size.

2.Q.2

The number of predictors p is extremely large, and the sample size n is small:

Gyu-Ho's Answer: Flexible is worse, because complex functions can be overfitting the data.

The flexible model will cause overfitting due to our small sample size.

2.Q.3

The relationship between the predictors and response is highly non-linear:

Gyu-Ho's Answer: Flexible is better, because simple linear regression would have high bias.

A flexible model will be necessary to find the nonlinear effect.

2.Q.4

The variance of the error terms, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high:

Gyu-Ho's Answer: Flexible is worse, because flexible methods would be fitting the irreducible error terms.

A flexible model will cause us to fit too much of the noise in the problem.