Chapter 2 Overview of Statistical Learning

1. Suppose we observe a quantitative response and p different predictors . Assume there is some relationship between and :

Here f is some fixed but unknown function of , and is a random error term independent of and has mean zero.

1. For any estimate of , we have

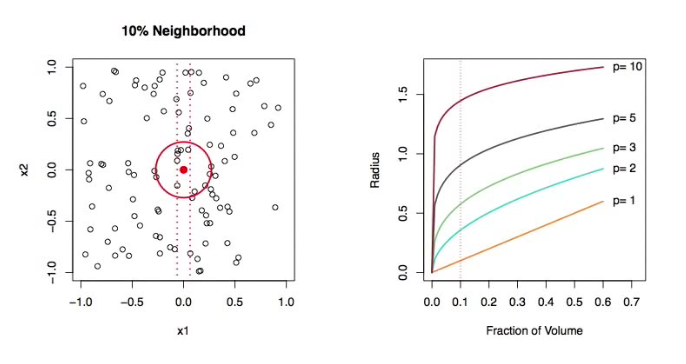
where is reducible error since we can potentially improve , and is irreducible error.

1. How to estimate
   1. Option 1: Estimate .

However, we may have few if any data point at exactly, and thus cannot compute it.

* 1. Option 2: Relax the definition and let , where is some neighborhood of x.

However, we may fall into curse of dimensionality: the nearest neighbors tend to be far away in high dimensions. For example, to include 10% of all data in 1-D uniformly distributed data, the fraction of length is only 10%, and thus radius is 5%; but in 2-D uniformly distributed data, the radius of the circle with area 0.1 is 18%.



* 1. Parametric and structured models, e.g., linear models. Need to consider a few trade-offs: (a) prediction accuracy versus interpretability; (b) good fit versus overfit/underfit.

1. How to assess model accuracy: mean square error on training data and test data
2. Bias-Variance trade-off

For a given value ,

where:

* is the expected test MSE at
* is the variance of if we estimated using a different training dataset
* is the bias between the model and the real-life problem

As a general rule, a more flexible model tends to have lower bias and higher variance.

Chapter 3 Linear Regression

1. Multiple regression may give a different result from if we run simple regression for each predictor separately. Reason: multiple regression coefficients represent the average change of associated with while holding other fixed; in contrast, simple regression coefficients ignore other predictors.

E.g., a simple regression of shark attacks on ice cream sales would suggest a positive relationship. However, there is no direct relationship: high temperatures cause more people to visit the beach, and this more ice cream sales and more shark attacks.

1. Important questions in linear regression
   1. Is at least one of predictors useful in predicting the response?

Answer: F test.

* 1. Do all predictors help to explain , or is only a subset of predictors useful?

Answer: Variable selection problems are extensively studied in Chapter 6. Some classical approaches are: (a) use some statistics to compare the models, such as

(b) when there are too many possible models that we cannot compare all of them, use forward/backward/mixed selection.

* 1. How well does the model fit the data?

Answer: Use or visualization. Note: in multiple regression,

* 1. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Answer: There are three sorts of uncertainty for a prediction: (1) is estimate of , reducible error; (2) model bias: the true model may not be linear, while we use a linear model; (3) random error, irreducible error.

1. Confidence interval vs. prediction interval: confidence interval is used to quantify the uncertainty of mean of y given x; prediction interval is used to quantify the uncertainty of y given x. Prediction interval is wider than confidence interval.
2. How to choose the subset the predictors in the regression model?
   1. Forward selection: begin with null model with only an intercept and no predictors. Each time add a variable that results in the lowest RSS, until some stopping rule is satisfied.
   2. Backward selection: begin with all variables in the model. Each time remove a variable with the largest p-value, until some stopping rule is satisfied.
3. The hierarchy principle for interaction term: if we include an interaction in a model, we should also include the main effects, even if the p-values associated with main effects are not significant.

Rationale: interaction term measures the change of a main effect when another main affect changes. The interaction term would be hard to interpret without main effects in the model.

1. Potential problems of linear regression: (not included in the lectures; see text section 3.3.3; to be added later)
   1. Non-linearity relationship

Solution: plot residual versus x. If non-linear relationship is detected, then add the non-linear term in the model.

* 1. Correlation of error: If the error terms are correlated, we will underestimate the variance term, causing us to erroneously conclude that a parameter is significant.

Solution: plot residuals versus t.

* 1. Non-constant variance of error terms

Solution: plot residuals versus fitted values; transform Y when necessary.

* 1. Outliers: an outlier is a point for which is far from the predicted value.

Solution: plot studentized residuals to detect outliers; outlier handling will be a case by case problem

* 1. High-leverage points: have an unusual value for x

Solution: visualization for simple regression; leverage statistic for multiple regression

* 1. Collinearity

Solution: correlation matrix of predictors (only for collinearity between pair of predictors); VIF

Chapter 4 Classification

For now have no thorough understanding of classification. To review in the future.

1. Logistic regression
2. Logistic regression with more than 2 classes

There is a exponential linear function for each class. Only degrees of freedom.

Chapter 5 Resampling Methods

1. A common mistake in cross validation: It is wrong to do model selection/variable selection on the whole dataset, and then further select models using cross validation. The reason is you already use information of whole dataset to do part of the model selection, and the test error would be underestimated.

Chapter 6 Linear Model Selection and Regularization

Three classes of model selection methods: subset selection, shrinkage, dimension reduction.

1. Subset selection
   1. Best subset selection, forward stepwise, backward stepwise
   2. Among the models with same number of predictors, we can use RSS or to decide which model is better.
   3. For models with different number of predictors, we have to either (a) make an adjustment to the training error, including ; or (b) use validation to directly estimate test error.
   4. In case of linear model with Gaussian errors, is equivalent to , and they will select the same model; tends to select a simpler model since the penalty term in is larger.
   5. One advantage of validation is, it does not require an estimate of
2. Shrinkage methods
   1. Ridge regression
   2. Lasso regression
   3. Scale equivalent: Least square regression is scale equivalent: if we scale a predictor from dollar to thousand dollar, the model will be equivalent; however, this is not the case for ridge and lasso. Therefore, we need to standardize the predictor by dividing by their standard deviation.
   4. The improvement over least square is from bias-variance tradeoff: decreased variance but increased bias.
   5. How to choose : cross validation
3. Dimension reduction methods
   1. Principal component regression: use principal components of to do regression. Note that the choice of principal components is independent of , so it may not choose the components that are most related to .
   2. Partial least squares.
   3. Before using PCR or PLS, we generally need to standardize the predictors. However, when all predictors are measured in the same unit, we can choose not to standardize.

Chapter 7 Moving Beyond Linearity

Chapter 8 Tree-Based Methods

1. Tree pruning: first build a large tree, then prune the tree to remove some nodes.
2. Bagging: bootstrap training dataset and build a tree from each sample, then use average value from these trees. When using bagging, we can build a big tree and do not need to prune. The reason is pruning aims at reducing variance, but bagging can reduce variance without pruning (pruning increases bias).
3. Random forest: similar to bagging that it builds many trees from bootstrap samples; but the difference is: at each split, we are only allowed to use a random sample of predictors, rather than the full set of predictors. We typically choose
4. Boosting.

Chapter 9 Support Vector Classifiers

1. Maximal margin classifier: find a hyperplane that perfectly separates the training data; only works when a separating hyperplane exists; not robust to outliers.
2. Support vector classifiers: find a hyperplane and a soft margin that correctly classify most points, but allow some points on the wrong side of margin/hyperplane, to achieve greater robustness and lower variance.
3. Support vector machines: the support vector classifiers can only generate linear boundary, we hope to extend the feature space. However, there can be infinite features, so we want to extend the feature space with efficient computations. By using different kernels, we can generate different support vector machines.

For example, if kernel is inner product, the SVM will be the support vector classifier with linear boundary; if kernel is radial kernel, ,the SVM will classify according to nearest points.

1. (From exercise) The beauty of the "kernel trick" is that, even if there is an infinite-dimensional basis, we need only look at the n^2 inner products between training data points.

Chapter 10 Unsupervised Learning