Lecture 28: Review

Reading: Relevant chapters in ISLR.

STATS 202: Data mining and analysis

Jelena Markovic
slide credits: Sergio Bacallado
October 23, 2018

Unsupervised learning

- In unsupervised learning, all the variables are on equal standing, no such thing as an input and response.
- **Two sets of methods:**
  1. PCA: find the main directions of variation in the data
  2. Clustering: find meaningful groups of samples
     - Hierarchical clustering (single, complete, or average linkage).
     - K-means clustering.

PCA

1. Find the linear combination of variables

$$\theta_{11}X_1 + \theta_{12}X_2 + \cdots + \theta_{1p}X_p$$

with $\sum_i \theta_{1i}^2 = 1$, which has the largest variance.

PCA

1. Find the linear combination of variables

$$\theta_{11}X_1 + \theta_{12}X_2 + \cdots + \theta_{1p}X_p$$

   with $\sum_i \theta_{1i}^2 = 1$, which has the largest variance.

2. Find the linear combination of variables

$$\theta_{21}X_1 + \theta_{22}X_2 + \cdots + \theta_{2p}X_p$$

   with $\sum_i \theta_{2i}^2 = 1$ and $\theta_1 \perp \theta_2$, which has the largest variance.

PCA

1. Find the linear combination of variables

$$\theta_{11}X_1 + \theta_{12}X_2 + \cdots + \theta_{1p}X_p$$

   with $\sum_i \theta_{1i}^2 = 1$, which has the largest variance.
2. Find the linear combination of variables

$$\theta_{21}X_1 + \theta_{22}X_2 + \cdots + \theta_{2p}X_p$$

   with $\sum_i \theta_{2i}^2 = 1$ and $\theta_1 \perp \theta_2$, which has the largest variance.
3. ...

PCA

Some questions:

- What are the loadings?

PCA

Some questions:

- What are the loadings?
- What are score variables?

PCA

Some questions:

- What are the loadings?
- What are score variables?
- What is a biplot, how is it interpreted?

PCA

Some questions:

- What are the loadings?
- What are score variables?
- What is a biplot, how is it interpreted?
- What is the proportion of variance explained? A scree plot?

PCA

Some questions:

- What are the loadings?
- What are score variables?
- What is a biplot, how is it interpreted?
- What is the proportion of variance explained? A scree plot?
- What is the effect of rescaling variables?

K-means clustering

- The number of clusters is fixed at K.

K-means clustering

- The number of clusters is fixed at K.
- Goal is to minimize the average distance of a point to the average of its cluster.

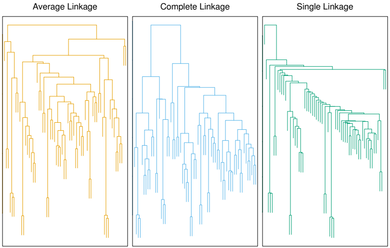
K-means clustering

- The number of clusters is fixed at K.
- Goal is to minimize the average distance of a point to the average of its cluster.
- The algorithm starts from some assignment, and is guaranteed to decrease this average distance.

K-means clustering

- The number of clusters is fixed at K.
- Goal is to minimize the average distance of a point to the average of its cluster.
- The algorithm starts from some assignment, and is guaranteed to decrease this average distance.
- This find a local minimum, not necessarily a global minimum, so we typically repeat the algorithm from many different random starting points.

Hierarchical clustering



- Agglomerative algorithm produces a *dendrogram*.
- At each step we join the two clusters that are "closest":
  - **Complete:** distance between clusters is maximal distance between any pair of points.
  - **Single:** distance between clusters is minimal distance.
  - **Average:** distance between clusters is the average distance.
- Height of a branching point = distance between clusters joined.

Supervised learning

Now, we have a response variable y_i associated to each vector of predictors x_i.

Supervised learning

Now, we have a response variable y_i associated to each vector of predictors x_i.

Two classes of problem:

- Regression: y_i is numerical

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Classification: y_i is categorical

$$0 - 1 \text{ loss} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(y_i \neq \hat{y}_i).$$

Training vs. test error

Both the MSE for regression, and the 0-1 loss for classification can be computed:

1. On the training data.
2. On an independent test set.

Training vs. test error

Both the MSE for regression, and the 0-1 loss for classification can be computed:

1. On the training data.
2. On an independent test set.

We want to minimize the error on a very large test set which is sampled from the same process as the training data. This is called the *test error*.

Bias-variance decomposition

Consider a regression method, which given some data $(x_1, y_1), \ldots, (x_n, y_n)$ outputs a prediction $\hat{f}(x)$ for the regression function.

Bias-variance decomposition

Consider a regression method, which given some data $(x_1, y_1), \ldots, (x_n, y_n)$ outputs a prediction $\hat{f}(x)$ for the regression function.

If we think of the training data as coming from some distribution, then the function $\hat{f}$ can be considered a random variable as well.

Bias-variance decomposition

Consider a regression method, which given some data $(x_1, y_1), \ldots, (x_n, y_n)$ outputs a prediction $\hat{f}(x)$ for the regression function.

If we think of the training data as coming from some distribution, then the function $\hat{f}$ can be considered a random variable as well.

The expected test MSE of $\hat{f}$ has the following decomposition for any fixed x:

$$E[(y - \hat{f}(x))^2] = \underbrace{E[(\hat{f}(x) - E\hat{f}(x))^2]}_{\text{Var}(\hat{f}(x)) > 0} + \underbrace{(E[\hat{f}(x) - f(x)])^2}_{\text{Square bias of } \hat{f}(x). > 0} + \text{Var}(\epsilon)$$

Variance: Increases with the flexibility of the model
Bias: Decreases as the flexibility of the model increases

Regression methods

- Nearest neighbors regression
- Multiple linear regression

Classification methods

- Nearest neighbors classification
- Logistic regression
- LDA and QDA

Self testing questions

For each of the regression and classification methods:

1. What are we trying to optimize?

Self testing questions

For each of the regression and classification methods:

1. What are we trying to optimize?
2. What does the fitting algorithm consist of, roughly?

Self testing questions

For each of the regression and classification methods:

1. What are we trying to optimize?
2. What does the fitting algorithm consist of, roughly?
3. What are the tuning parameters, if any?

Self testing questions

For each of the regression and classification methods:

1. What are we trying to optimize?
2. What does the fitting algorithm consist of, roughly?
3. What are the tuning parameters, if any?
4. How is the method related to other methods, mathematically and in terms of bias, variance?

Self testing questions

For each of the regression and classification methods:

1. What are we trying to optimize?
2. What does the fitting algorithm consist of, roughly?
3. What are the tuning parameters, if any?
4. How is the method related to other methods, mathematically and in terms of bias, variance?
5. How does rescaling or transforming the variables affect the method?

Self testing questions

For each of the regression and classification methods:

1. What are we trying to optimize?
2. What does the fitting algorithm consist of, roughly?
3. What are the tuning parameters, if any?
4. How is the method related to other methods, mathematically and in terms of bias, variance?
5. How does rescaling or transforming the variables affect the method?
6. In what situations does this method work well? What are its limitations?

Evaluating a classification method

We have talked about the 0-1 loss:

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}(y_i \neq \hat{y}_i).$$

It is possible to make the wrong prediction for some classes more often than others. The 0-1 loss doesn't tell you anything about this.

Evaluating a classification method

We have talked about the 0-1 loss:

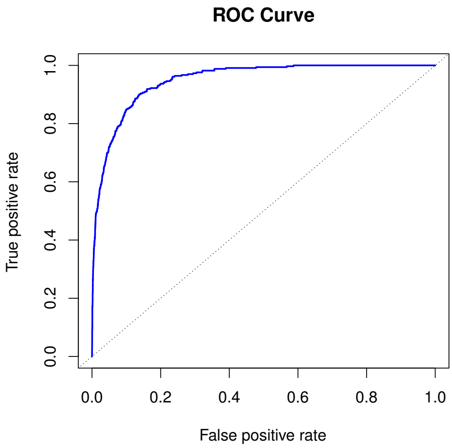
$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}(y_i \neq \hat{y}_i).$$

It is possible to make the wrong prediction for some classes more often than others. The 0-1 loss doesn't tell you anything about this.

A much more informative summary of the error is a **confusion matrix**:

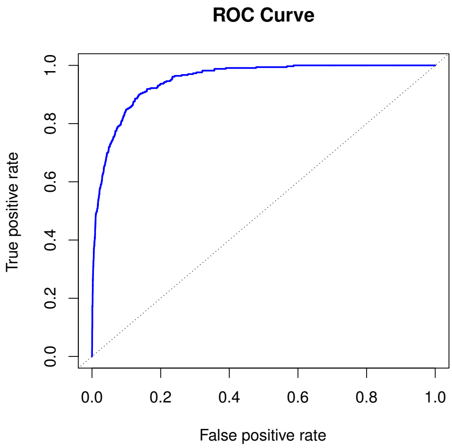
		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

The ROC curve



- Displays the performance of the method for any choice of threshold.

The ROC curve



- Displays the performance of the method for any choice of threshold.
- The area under the curve (AUC) measures the quality of the classifier:
  - 0.5 is the AUC for a random classifier
  - The closer AUC is to 1, the better.

How do we estimate the test error?

- Our main technique is cross-validation.

How do we estimate the test error?

- Our main technique is cross-validation.
- Different approaches:
  1. **Validation set:** Split the data in two parts, train the model on one subset, and compute the test error on the other.

How do we estimate the test error?

- Our main technique is cross-validation.
- Different approaches:
  1. **Validation set:** Split the data in two parts, train the model on one subset, and compute the test error on the other.
  2. k-**fold:** Split the data into k subsets. Average the test errors computed using each subset as a validation set.

How do we estimate the test error?

- Our main technique is cross-validation.
- Different approaches:
  1. **Validation set:** Split the data in two parts, train the model on one subset, and compute the test error on the other.
  2. k-**fold:** Split the data into k subsets. Average the test errors computed using each subset as a validation set.
  3. **LOOCV:** k-fold cross validation with $k = n$.

How do we estimate the test error?

- Our main technique is cross-validation.
- Different approaches:
  1. **Validation set:** Split the data in two parts, train the model on one subset, and compute the test error on the other.
  2. k-**fold:** Split the data into k subsets. Average the test errors computed using each subset as a validation set.
  3. **LOOCV:** k-fold cross validation with $k = n$.
- No approach is superior to all others.

How do we estimate the test error?

- Our main technique is cross-validation.
- Different approaches:
  1. **Validation set:** Split the data in two parts, train the model on one subset, and compute the test error on the other.
  2. k-**fold:** Split the data into k subsets. Average the test errors computed using each subset as a validation set.
  3. **LOOCV:** k-fold cross validation with $k = n$.
- No approach is superior to all others.
- What are the main differences? How do the bias and variance of the test error estimates compare? Which methods depend on the random seed?

The Bootstrap

- **Main idea:** If we have enough data, the empirical distribution is similar to the actual distribution of the data.
- Resampling with replacement allows us to obtain datasets mimicing how original data was selected.