Lecture Outline

ROC Curves

k-NN Revisited

Dealing with Missing Data

Types of Missingness

Imputation Methods

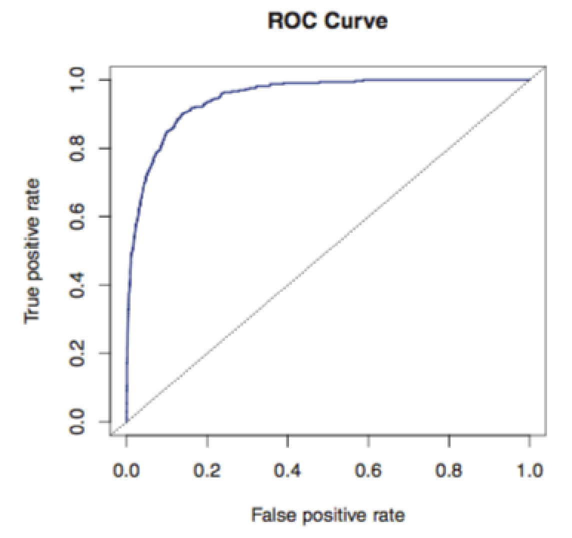
The ROC curve illustrates the trade-off for all possible thresholds chosen for the two types of error (or correct classification).

The vertical axis displays the true positive predictive value and the horizontal axis depicts the true negative predictive value.

What is the shape of an ideal ROC curve?

See next slide for an example.

ROC Curve Example



k-NN for Classification

How can we modify the k-NN approach for classification?

How can we modify the k-NN approach for classification?

The approach here is the same as for k-NN regression: use the other available observations that are most similar to the observation we are trying to predict (classify into a group) based on the predictors at hand.

How do we classify which category a specific observation should be in based on its nearest neighbors?

The category that shows up the most among the nearest neighbors.

k-NN with Multiple Predictors

How could we extend k-NN (both regression and classification) when there are multiple predictors?

We would need to define a measure of distance for observations in order to which are the most similar to the observation we are trying to predict.

Euclidean distance is a good option. To measure the distance of a new observation, $\mathbf{x}_0$ from each observation in the data set, $\mathbf{x}_i$:

$$D^2(\mathbf{x}_i, \mathbf{x}_0) = \sum_{j=1}^{P} (x_{i,j} - x_{0,j})^2$$

Dealing with Missing Data

What is missing data?

Often times when data is collected, there are some missing values apparent in the dataset. This leads to a few questions to consider:

1. How does this show up in pandas?
2. How does pandas and sklearn handle these NaNs?
3. How does this effect our modeling?

Naively handling missingness

What is the simplest way to handle missing data?

1. Impute the mean (if quantitative) or most common class (if categorical) for all missing values.
2. How does pandas and sklearn handle these NaNs?

What are some consequences in handling missingness in this fashion?

Missing data can arise from various places in data:

- A survey was conducted and values were just randomly missed when being entered in the computer.
- A respondent chooses not to respond to a question like ‘Have you ever done cocaine?’.
- You decide to start collecting a new variable (like Mac vs. PC) partway through the data collection of a study.
- You want to measure the speed of meteors, and some observations are just ’too quick’ to be measured properly.

The source of missing values in data can lead to the major types of missingness:

Types of Missingness

There are 3 major types of missingness to be concerned about:

1. **Missing Completely at Random (MCAR)** - the probability of missingness in a variable is the same for all units. Like randomly poking holes in a data set.
2. **Missing at Random (MAR)** - the probability of missingness in a variable depends only on available information (in other predictors).
3. **Missing Not at Random (MNAR)** - the probability of missingness depends on information that has not been recorded and this information also predicts the missing values.

What are examples of each these 3 types?

Missing completely at random (MCAR)

Missing Completely at Random is the best case scenario, and the easiest to handle:

- Examples: a coin is flipped to determine whether an entry is removed. Or when values were just randomly missed when being entered in the computer.
- Effect if you ignore: there is no effect on inferences (estimates of beta).
- How to handle: lots of options, but best to impute (more on next slide)

Missing at random (MAR)

Missing at random is still a case that can be handled.

- Example(s): men and women respond to the question have you ever felt harassed at work? at different rates (and may be harassed at different rates).
- Effect if you ignore: inferences are biased (estimates of beta) and predictions are usually worsened.
- How to handle: use the information in the other predictors to build a model and 'impute' a value for the missing entry.

Key: we can fix any biases by modeling and imputing the missing values based on what is observed!

Missing Not at Random (MNAR)

Missing Not at Random is the worst case scenario, and impossible to handle:

- Example(s): patients drop out of a study because they experience some really bad side effect that was not measured. Or cheaters are less likely to respond when asked if you've ever cheated.
- Effect if you ignore: there is no effect on inferences (estimates of beta) or predictions.
- How to handle: you can 'improve' things by dealing with it like it is MAR, but you [likely] may never completely fix the bias.

What type of missingness is present?

Can you ever tell based on your data what type of missingness is actually present?

Since we asked the question, the answer must be no. It generally cannot be determined whether data really are missing at random, or whether the missingness depends on unobserved predictors or the missing data themselves. The problem is that these potential lurking variables are unobserved (by definition) and so can never be completely ruled out.

In practice, a model with as many predictors as possible is used so that the 'missing at random' assumption is reasonable.

When encountering missing data, the approach to handling it depends on:

1. whether the missing values are in the response or in the predictors. Generally speaking, it is much easier to handle missingness in predictors.
2. whether the variable is quantitative or categorical.
3. how much missingness is present in the variable. If there is too much missingness, you may be doing more damage than good.

Generally speaking, it is a good idea to attempt to **impute** (or 'fill in') entries for missing values in a variable (assuming your method of imputation is a good one).

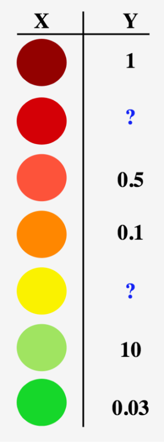
There are several different approaches to imputing missing values:

1. Plug in the mean (quantitative) or most common class (categorical) for all missing values in a variable.
2. Create a new variable that is an indicator of missingness, and include it in any model to predict the response (also plug in zero or the mean in the actual variable).
3. Hot deck imputation: for each missing entry, randomly select an observed entry in the variable and plug it in.
4. Model the imputation: plug in predicted values ($\hat{y}$) from a model based on the other observed predictors.
5. Model the imputation with uncertainty: plug in predicted values plus randomness ($\hat{y} + \varepsilon$) from a model based on the other observed predictors.

What are the advantages and disadvantages of each approach?

Schematic: imputation through modeling

How do we use models to fill in missing data?

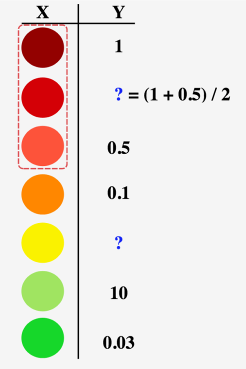


Schematic: imputation through modeling

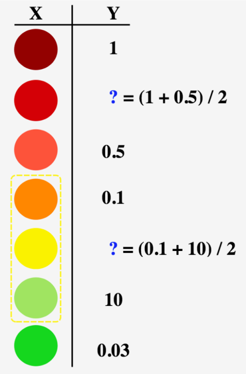
How do we use models to fill in missing data?

Schematic: imputation through modeling

How do we use models to fill in missing data? Using kNN for $k = 2$?



How do we use models to fill in missing data? Using kNN for $k = 2$?



Schematic: imputation through modeling

How do we use models to fill in missing data? Using linear regression?

Schematic: imputation through modeling

How do we use models to fill in missing data? Using linear regression?

Where m and b are computed from the observations (rows) that do not have missingness (we should call them $b = \beta_0$ and $m = \beta_1$).

Recall the probabilistic model in linear regression:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$. How can we take advantage of this model to impute with uncertainty?

It's a 3 step process:

1. Fit a model to predict the predictor variable with missingness from all the other predictors.
2. Predict the missing values from the model in the previous part.
3. Add in a measure of uncertainty to this prediction by randomly sampling from a $N(0, \hat{\sigma}^2)$ distribution, where $\hat{\sigma}^2$ is the mean square error (MSE) from the model.

Imputation through modeling with uncertainty: k-NN regression

How can we use k-NN regression to impute values that mimic the error in our observations?

How can we use k-NN regression to impute values that mimic the error in our observations?

Two ways:

1. Use $k = 1$.
2. Use any other k, but randomly select from the nearest neighbors in $\mathcal{N}_0$. This can be done with equal probability or with some weighting (inverse to the distance measure used).