Introduction to Statistical Learning

Gaurav Sood

Spring 2015

Two paradigms of learning

Two paradigms of learning

– Supervised Learning

Two paradigms of learning

- Supervised Learning
  - When getting labels (predictions) is expensive

Two paradigms of learning

- Supervised Learning
  - When getting labels (predictions) is expensive
  - Get labels for a small set of data

Two paradigms of learning

- Supervised Learning
  - When getting labels (predictions) is expensive
  - Get labels for a small set of data
  - Estimate relationship between Y and X

Two paradigms of learning

- Supervised Learning
  - When getting labels (predictions) is expensive
  - Get labels for a small set of data
  - Estimate relationship between Y and X
  - Predict labels of unseen data

Two paradigms of learning

- Supervised Learning
  - When getting labels (predictions) is expensive
  - Get labels for a small set of data
  - Estimate relationship between Y and X
  - Predict labels of unseen data
  - Labels and cost function *supervise* dimension reduction

Two paradigms of learning

- Supervised Learning
  - When getting labels (predictions) is expensive
  - Get labels for a small set of data
  - Estimate relationship between Y and X
  - Predict labels of unseen data
  - Labels and cost function *supervise* dimension reduction
- Unsupervised Learning

Two paradigms of learning

- Supervised Learning
  - When getting labels (predictions) is expensive
  - Get labels for a small set of data
  - Estimate relationship between Y and X
  - Predict labels of unseen data
  - Labels and cost function *supervise* dimension reduction
- Unsupervised Learning
  - Find vectors similar to each other, maximize differences across

Two paradigms of learning

- Supervised Learning
  - When getting labels (predictions) is expensive
  - Get labels for a small set of data
  - Estimate relationship between Y and X
  - Predict labels of unseen data
  - Labels and cost function *supervise* dimension reduction
- Unsupervised Learning
  - Find vectors similar to each other, maximize differences across
  - Find rows similar to each other, maximize differences across

How to learn from data?

- $Y = f(X) + \epsilon$

How to learn from data?

- $Y = f(X) + \epsilon$
- How do we estimate $f(X)$?

How to learn from data?

- $Y = f(X) + \epsilon$
- How do we estimate $f(X)$?
- If similar x, similar y

How to learn from data?

- $Y = f(X) + \epsilon$
- How do we estimate $f(X)$?
- If similar x, similar y
- Function: value of y same as that of the nearest neighbor

How to learn from data?

- $Y = f(X) + \epsilon$
- How do we estimate $f(X)$?
- If similar x, similar y
- Function: value of y same as that of the nearest neighbor
- Question and a concern

How to learn from data?

- $Y = f(X) + \epsilon$
- How do we estimate $f(X)$?
- If similar x, similar y
- Function: value of y same as that of the nearest neighbor
- Question and a concern
  - What do we mean by nearest?

How to learn from data?

- $Y = f(X) + \epsilon$
- How do we estimate $f(X)$?
- If similar x, similar y
- Function: value of y same as that of the nearest neighbor
- Question and a concern
  - What do we mean by nearest?
  - Wouldn't it depend on what x are observed?

What do we mean by nearest?

What do we mean by nearest?

- Euclidean distance: If p and q are two n dimensional vectors

$$d_e(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

What do we mean by nearest?

- Euclidean distance: If p and q are two n dimensional vectors
  $d_e(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$
- Some issues:

What do we mean by nearest?

- Euclidean distance: If p and q are two n dimensional vectors

$$d_e(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

- Some issues:
  - Not all features on the same scale

What do we mean by nearest?

- Euclidean distance: If p and q are two n dimensional vectors
  $$d_e(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$
- Some issues:
  - Not all features on the same scale
  - Features may be correlated with each other

What do we mean by nearest?

- Euclidean distance: If p and q are two n dimensional vectors
  $d_e(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$
- Some issues:
  - Not all features on the same scale
  - Features may be correlated with each other
- Mahalanobis distance between two vectors:
  Say S is the covariance matrix
  $d_m(\vec{p},\vec{q}) = \sqrt{(\vec{p}-\vec{q})'S^{-1}(\vec{p}-\vec{q})}$

What do we mean by nearest?

- Euclidean distance: If p and q are two n dimensional vectors
  $d_e(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$
- Some issues:
  - Not all features on the same scale
  - Features may be correlated with each other
- Mahalanobis distance between two vectors:
  Say S is the covariance matrix
  $d_m(\vec{p},\vec{q}) = \sqrt{(\vec{p}-\vec{q})'S^{-1}(\vec{p}-\vec{q})}$
- For Boolean, Jaccard distance:
  $d_j(p,q) = \frac{|p \cup q| - |p \cap q|}{|p \cup q|}$
  Applications: Recommender systems, finding similar

Learning from data

– $Y = f(X) + e$

Learning from data

- $Y = f(X) + e$
- Problem formulated as a regression function

Learning from data

- $Y = f(X) + e$
- Problem formulated as a regression function

Learning from data

- $Y = f(X) + e$
- Problem formulated as a regression function
  - $E(Y|X)$

Learning from data

- $Y = f(X) + e$
- Problem formulated as a regression function
  - $E(Y|X)$
  - $f(x) = E(Y|x = x)$

Learning from data

- $Y = f(X) + e$
- Problem formulated as a regression function
  - $E(Y|X)$
  - $f(x) = E(Y|x = x)$
- Nearest neighbour averaging
  $\hat{f}(x) = E[Y|X \in N(x)]$

Learning from data

- $Y = f(X) + e$
- Problem formulated as a regression function
  - $E(Y|X)$
  - $f(x) = E(Y|x = x)$
- Nearest neighbour averaging
  $\hat{f}(x) = E[Y|X \in N(x)]$
- Great when small p, large N

Curse of Dimensionality

– $\hat{f}(x) = E[Y|X \in N(x)]$

Curse of Dimensionality

- $\hat{f}(x) = E[Y|X \in N(x)]$
- Define neighborhood too tightly, nothing in there.

Curse of Dimensionality

- $\hat{f}(x) = E[Y|X \in N(x)]$
- Define neighborhood too tightly, nothing in there.
- If we expand it, nearest neighbors can be far away

Curse of Dimensionality

- $\hat{f}(x) = E[Y|X \in N(x)]$
- Define neighborhood too tightly, nothing in there.
- If we expand it, nearest neighbors can be far away
- How do we solve the problem?

Curse of Dimensionality

- $\hat{f}(x) = E[Y|X \in N(x)]$
- Define neighborhood too tightly, nothing in there.
- If we expand it, nearest neighbors can be far away
- How do we solve the problem?
- One way: parametric and structural models $f(x) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ...\beta_p * X_p$

How to expand the linear model?

- Simple polynomial transformations

How to expand the linear model?

- Simple polynomial transformations
- Interactions

How to expand the linear model?

- Simple polynomial transformations
- Interactions
- Step functions. E.g. transform education into k dummy variables.

How to expand the linear model?

- Simple polynomial transformations
- Interactions
- Step functions. E.g. transform education into k dummy variables.
- More complicated basis functions:

How to expand the linear model?

- Simple polynomial transformations
- Interactions
- Step functions. E.g. transform education into k dummy variables.
- More complicated basis functions:
  - Piecewise polynomial: takes differential polynomial for different regions (split by knots)

How to expand the linear model?

- Simple polynomial transformations
- Interactions
- Step functions. E.g. transform education into k dummy variables.
- More complicated basis functions:
  - Piecewise polynomial: takes differential polynomial for different regions (split by knots)
  - Add constraint that there are no abrupt changes across regions.

How to expand the linear model?

- Simple polynomial transformations
- Interactions
- Step functions. E.g. transform education into k dummy variables.
- More complicated basis functions:
  - Piecewise polynomial: takes differential polynomial for different regions (split by knots)
  - Add constraint that there are no abrupt changes across regions.
  - Add constraint that first and second derivates are the same. (For cubic splines.)

How to expand the linear model?

- Simple polynomial transformations
- Interactions
- Step functions. E.g. transform education into k dummy variables.
- More complicated basis functions:
  - Piecewise polynomial: takes differential polynomial for different regions (split by knots)
  - Add constraint that there are no abrupt changes across regions.
  - Add constraint that first and second derivates are the same. (For cubic splines.)
  - Add boundary constraint: linear before 1st knot or after last knot. (Natural spline.)

Error

Error

- Reducible error:
  $Var(\hat{f}) + [\text{Bias}(E[\hat{f}(x) - f(x)])]^2$

Error

- Reducible error:
  $Var(\hat{f}) + [\text{Bias}(E[\hat{f}(x) - f(x)])]^2$
- Bias-Variance tradeoff

Error

- Reducible error:
  $Var(\hat{f}) + [\text{Bias}(E[\hat{f}(x) - f(x)])]^2$
- Bias-Variance tradeoff
- As flexibility increases, $Var(\hat{f})$ increases
  Model goes after each wrinkle in the training data

Error

- Reducible error:
  $Var(\hat{f}) + [\mathrm{Bias}(E[\hat{f}(x) - f(x)])]^2$
- Bias-Variance tradeoff
- As flexibility increases, $Var(\hat{f})$ increases
  Model goes after each wrinkle in the training data
- Say we have estimated the ideal function $\hat{f}(X)$

Error

- Reducible error:
  $Var(\hat{f}) + [\text{Bias}(E[\hat{f}(x) - f(x)])]^2$
- Bias-Variance tradeoff
- As flexibility increases, $Var(\hat{f})$ increases
  Model goes after each wrinkle in the training data
- Say we have estimated the ideal function $\hat{f}(X)$
- Ideal w.r.t a loss function, e.g., average squared error: $(Y - \hat{Y})^2$

Error

- Reducible error:
  $Var(\hat{f}) + [\text{Bias}(E[\hat{f}(x) - f(x)])]^2$
- Bias-Variance tradeoff
- As flexibility increases, $Var(\hat{f})$ increases
  Model goes after each wrinkle in the training data
- Say we have estimated the ideal function $\hat{f}(X)$
- Ideal w.r.t a loss function, e.g., average squared error: $(Y - \hat{Y})^2$
- Ideal still leaves some error (irreducible error):
  - $\epsilon = Y - \hat{f}(x)$

Error

- Reducible error:
  $Var(\hat{f}) + [\text{Bias}(E[\hat{f}(x) - f(x)])]^2$
- Bias-Variance tradeoff
- As flexibility increases, $Var(\hat{f})$ increases
  Model goes after each wrinkle in the training data
- Say we have estimated the ideal function $\hat{f}(X)$
- Ideal w.r.t a loss function, e.g., average squared error: $(Y - \hat{Y})^2$
- Ideal still leaves some error (irreducible error):
  - $\epsilon = Y - \hat{f}(x)$
  - $E[(Y - \hat{f}(X)^2|X = x)] = (f(x) - \hat{f}(X))^2 + Var(\epsilon)$

Evaluating Models

Evaluating Models

- Deviance
  - Deviance $\propto -$Log-Likelihood

Evaluating Models

- Deviance
  - Deviance $\propto -$Log-Likelihood
  - Likelihood: $p(y_1|x_1)xp(y_2|x_2)x...xp(y_n|x_n)$

Evaluating Models

- Deviance
  - Deviance $\propto -$Log-Likelihood
  - Likelihood: $p(y_1|x_1)xp(y_2|x_2)x...xp(y_n|x_n)$
  - $\hat{\beta}$ maximize Likelihood (or minimize Deviance)

Evaluating Models

- Deviance
  - Deviance $\propto -$Log-Likelihood
  - Likelihood: $p(y_1|x_1)xp(y_2|x_2)x...xp(y_n|x_n)$
  - $\hat{\beta}$ maximize Likelihood (or minimize Deviance)
  - $R^2 = \frac{\text{Deviance of Fitted Model}}{\text{Deviance of Null Model}}$

Evaluating Models

- Deviance
  - Deviance $\propto$ $-$Log-Likelihood
  - Likelihood: $p(y_1|x_1) x p(y_2|x_2) x ... x p(y_n|x_n)$
  - $\hat{\beta}$ maximize Likelihood (or minimize Deviance)
  - $R^2 = \frac{\text{Deviance of Fitted Model}}{\text{Deviance of Null Model}}$
- AIC

Evaluating Models

- Deviance
  - Deviance $\propto -$Log-Likelihood
  - Likelihood: $p(y_1|x_1) x p(y_2|x_2) x ... x p(y_n|x_n)$
  - $\hat{\beta}$ maximize Likelihood (or minimize Deviance)
  - $R^2 = \frac{\text{Deviance of Fitted Model}}{\text{Deviance of Null Model}}$
- AIC
  - $\text{Deviance} + 2 * \text{df}$

Evaluating Models

- Deviance
  - Deviance $\propto$ $-$Log-Likelihood
  - Likelihood: $p(y_1|x_1) x p(y_2|x_2) x ... x p(y_n|x_n)$
  - $\hat{\beta}$ maximize Likelihood (or minimize Deviance)
  - $R^2 = \frac{\text{Deviance of Fitted Model}}{\text{Deviance of Null Model}}$
- AIC
  - $\text{Deviance} + 2 * \text{df}$
  - In-sample - Out of sample Deviance $\sim 2 * \text{df}$

Evaluating Models

- Deviance
  - Deviance $\propto -$Log-Likelihood
  - Likelihood: $p(y_1|x_1) x p(y_2|x_2) x ... x p(y_n|x_n)$
  - $\hat{\beta}$ maximize Likelihood (or minimize Deviance)
  - $R^2 = \frac{\text{Deviance of Fitted Model}}{\text{Deviance of Null Model}}$
- AIC
  - Deviance $+ 2 * \text{df}$
  - In-sample - Out of sample Deviance $\sim 2 * \text{df}$
  - AIC $\sim$ Out of sample Deviance

Evaluating Models

- Deviance
  - Deviance $\propto -$Log-Likelihood
  - Likelihood: $p(y_1|x_1) x p(y_2|x_2) x ... x p(y_n|x_n)$
  - $\hat{\beta}$ maximize Likelihood (or minimize Deviance)
  - $R^2 = \frac{\text{Deviance of Fitted Model}}{\text{Deviance of Null Model}}$
- AIC
  - $\text{Deviance} + 2 * \text{df}$
  - In-sample - Out of sample Deviance $\sim 2 * \text{df}$
  - AIC $\sim$ Out of sample Deviance
  - AIC overfits in high dimensions (df $\sim$ n).

Evaluating Models

- Deviance
  - Deviance $\propto -$Log-Likelihood
  - Likelihood: $p(y_1|x_1) x p(y_2|x_2) x \ldots x p(y_n|x_n)$
  - $\hat{\beta}$ maximize Likelihood (or minimize Deviance)
  - $R^2 = \frac{\text{Deviance of Fitted Model}}{\text{Deviance of Null Model}}$
- AIC
  - $\text{Deviance} + 2 * \text{df}$
  - In-sample - Out of sample Deviance $\sim 2 * \text{df}$
  - AIC $\sim$ Out of sample Deviance
  - AIC overfits in high dimensions (df $\sim$ n).
  - AICc = $\text{Deviance} + 2 * \text{df} * \frac{n}{n - df - 1}$

Evaluating Models

- Deviance
  - Deviance $\propto -$Log-Likelihood
  - Likelihood: $p(y_1|x_1)xp(y_2|x_2)x...xp(y_n|x_n)$
  - $\hat{\beta}$ maximize Likelihood (or minimize Deviance)
  - $R^2 = \frac{\text{Deviance of Fitted Model}}{\text{Deviance of Null Model}}$
- AIC
  - $\text{Deviance} + 2 * \text{df}$
  - In-sample - Out of sample Deviance $\sim 2 * \text{df}$
  - AIC $\sim$ Out of sample Deviance
  - AIC overfits in high dimensions (df $\sim$ n).
  - AICc = $\text{Deviance} + 2 * \text{df} * \frac{n}{n-df-1}$
  - BIC = $\text{Deviance} + \text{df} * log(n)$
  - BIC underfits when large n.

Evaluating Classification Models

		Observed	
		true	false
Predicted	true	true positive	false positive
	false	false negative	true negative

Evaluating Classification Models

		Observed	
		true	false
Predicted	true	true positive	false positive
	false	false negative	true negative

– Confusion matrix, $c^2 - c$ total possible errors

Evaluating Classification Models

		Observed	
		true	false
Predicted	true	true positive	false positive
	false	false negative	true negative

– Confusion matrix, $c^2 - c$ total possible errors

– Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$

Evaluating Classification Models

		Observed	
		true	false
Predicted	true	true positive	false positive
	false	false negative	true negative

- Confusion matrix, $c^2 - c$ total possible errors
- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
- Error Rate: $\frac{FP+FN}{TP+TN+FP+FN}$

Evaluating Classification Models

		Observed	
		true	false
Predicted	true	true positive	false positive
	false	false negative	true negative

- Confusion matrix, $c^2 - c$ total possible errors
- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
- Error Rate: $\frac{FP+FN}{TP+TN+FP+FN}$
- Sensitivity, TPR: $\frac{TP}{TP+FN}$

Evaluating Classification Models

		Observed	
		true	false
Predicted	true	true positive	false positive
	false	false negative	true negative

- Confusion matrix, $c^2 - c$ total possible errors
- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
- Error Rate: $\frac{FP+FN}{TP+TN+FP+FN}$
- Sensitivity, TPR: $\frac{TP}{TP+FN}$
- Specificity, FPR: $\frac{TN}{FP+TN}$

Evaluating Classification Models

		Observed	
		true	false
Predicted	true	true positive	false positive
	false	false negative	true negative

- Confusion matrix, $c^2 - c$ total possible errors
- Accuracy: $\frac{TP+TN}{TP+TN+FP+FN}$
- Error Rate: $\frac{FP+FN}{TP+TN+FP+FN}$
- Sensitivity, TPR: $\frac{TP}{TP+FN}$
- Specificity, FPR: $\frac{TN}{FP+TN}$
- BER: $\frac{1}{2}(TPR + TNR)$

Evaluating Classification Models

– ROC: TPR Vs. FPR

Evaluating Classification Models

- ROC: TPR Vs. FPR
- Precision: fraction of retrieved instances that are relevant
  $\frac{TP}{TP+FP}$

Evaluating Classification Models

- ROC: TPR Vs. FPR
- Precision: fraction of retrieved instances that are relevant
  $\frac{TP}{TP+FP}$
- Recall: fraction of relevant instances that are retrieved
  $\frac{TP}{TP+FN}$

Evaluating Classification Models

- ROC: TPR Vs. FPR
- Precision: fraction of retrieved instances that are relevant
  $\frac{TP}{TP+FP}$
- Recall: fraction of relevant instances that are retrieved
  $\frac{TP}{TP+FN}$
- F_1: $2\frac{\text{precision*recall}}{\text{precision + recall}}$

Evaluating Classification Models

- ROC: TPR Vs. FPR
- Precision: fraction of retrieved instances that are relevant
  $\frac{TP}{TP+FP}$
- Recall: fraction of relevant instances that are retrieved
  $\frac{TP}{TP+FN}$
- F_1: $2\frac{\text{precision*recall}}{\text{precision + recall}}$
- F_β: $(1+\beta^2)\frac{\text{precision*recall}}{\beta^2\text{precision + recall}}$

Out of Sample Error

- Another way to assess model error

Out of Sample Error

- Another way to assess model error
- R^2 always increases with more covariates.

Out of Sample Error

- Another way to assess model error
- R^2 always increases with more covariates.
- Or: As model complexity increases, training error goes down.

Out of Sample Error

- Another way to assess model error
- R^2 always increases with more covariates.
- Or: As model complexity increases, training error goes down.
- But out of sample error goes down and then up.

Out of Sample Error

– Another way to assess model error

– R^2 always increases with more covariates.

– Or: As model complexity increases, training error goes down.

– But out of sample error goes down and then up.

– Out of sample R^2 can be worse than $\bar{y}$.

Out of Sample Error

- Another way to assess model error
- R^2 always increases with more covariates.
- Or: As model complexity increases, training error goes down.
- But out of sample error goes down and then up.
- Out of sample R^2 can be worse than $\bar{y}$.
- Use of out of sample error to prevent overfitting

Out of Sample Error

– Another way to assess model error

– R^2 always increases with more covariates.

– Or: As model complexity increases, training error goes down.

– But out of sample error goes down and then up.

– Out of sample R^2 can be worse than $\bar{y}$.

– Use of out of sample error to prevent overfitting

– Net prediction error on test set can vary a lot.

A Clarification

Error in thinking about errors

Error in thinking about errors

– In sciences, ‘data mining’ is a dirty ‘phrase’

Error in thinking about errors

– In sciences, ‘data mining’ is a dirty ‘phrase’

– Jealousy?

Error in thinking about errors

- In sciences, ‘data mining’ is a dirty ‘phrase’
- Jealousy?
- Evokes concerns about false positives . . .

Error in thinking about errors

- In sciences, ‘data mining’ is a dirty ‘phrase’
- Jealousy?
- Evokes concerns about false positives . . .
- But ‘mining’ is by definition ‘the extraction of valuable [stuff]’

Error in thinking about errors

- In sciences, ‘data mining’ is a dirty ‘phrase’
- Jealousy?
- Evokes concerns about false positives . . .
- But ‘mining’ is by definition ‘the extraction of valuable [stuff]’
- So – is more data worse?

Error in thinking about errors

- In sciences, ‘data mining’ is a dirty ‘phrase’
- Jealousy?
- Evokes concerns about false positives . . .
- But ‘mining’ is by definition ‘the extraction of valuable [stuff]’
- So – is more data worse?
- Not quite

Error in thinking about errors

– In sciences, ‘data mining’ is a dirty ‘phrase’

– Jealousy?

– Evokes concerns about false positives . . .

– But ‘mining’ is by definition ‘the extraction of valuable [stuff]’

– So – is more data worse?

– Not quite

– Larger n allows for more precise estimation of relationship

False Positives

– Significance testing:

False Positives

- Significance testing:
  - Say .05, 5% false positive rate

False Positives

- Significance testing:
  - Say .05, 5% false positive rate
  - Assume independence, 1 of 20 false positive

False Positives

- Significance testing:
  - Say .05, 5% false positive rate
  - Assume independence, 1 of 20 false positive
  - Say 100 vars, 5 true positives, all sig., 5% of 95 $\sim$ 5. So 50% false discovery rate.

False Positives

- Significance testing:
  - Say .05, 5% false positive rate
  - Assume independence, 1 of 20 false positive
  - Say 100 vars, 5 true positives, all sig., 5% of 95 $\sim$ 5. So 50% false discovery rate.
- Fixes:

False Positives

- Significance testing:
  - Say .05, 5% false positive rate
  - Assume independence, 1 of 20 false positive
  - Say 100 vars, 5 true positives, all sig., 5% of 95 $\sim$ 5. So 50% false discovery rate.
- Fixes:
  - Familywise error rate (Bonferroni)

False Positives

- Significance testing:
  - Say .05, 5% false positive rate
  - Assume independence, 1 of 20 false positive
  - Say 100 vars, 5 true positives, all sig., 5% of 95 $\sim$ 5. So 50% false discovery rate.
- Fixes:
  - Familywise error rate (Bonferroni)
  - Optimization can be done w.r.t. to cost of false positive and negative
    e.g. Increase cut-off marks in exams, Breast Cancer

False Positives

- Significance testing:
  - Say .05, 5% false positive rate
  - Assume independence, 1 of 20 false positive
  - Say 100 vars, 5 true positives, all sig., 5% of 95 $\sim$ 5. So 50% false discovery rate.
- Fixes:
  - Familywise error rate (Bonferroni)
  - Optimization can be done w.r.t. to cost of false positive and negative

    e.g. Increase cut-off marks in exams, Breast Cancer
  - False Discovery Rate

False Discovery Rate

$$\text{False discovery Proportion} = \frac{\#\text{ of FP}}{\#\text{ of Sig. Results}}$$

False Discovery Rate

$$\text{False discovery Proportion} = \frac{\#\text{ of FP}}{\#\text{ of Sig. Results}}$$

– Can't be known but we can produce cutoffs so $E(FDP) < q$

False Discovery Rate

$$\text{False discovery Proportion} = \frac{\#\text{ of FP}}{\#\text{ of Sig. Results}}$$

– Can't be known but we can produce cutoffs so $E(FDP) < q$

– Benjamini and Hochberg (1995):

False Discovery Rate

$$\text{False discovery Proportion} = \frac{\#\text{ of FP}}{\#\text{ of Sig. Results}}$$

- Can't be known but we can produce cutoffs so $E(FDP) < q$
- Benjamini and Hochberg (1995):
  - Rank the n p-values, smallest to largest, $p_1 \ldots p_n$.

False Discovery Rate

$$\text{False discovery Proportion} = \frac{\#\text{ of FP}}{\#\text{ of Sig. Results}}$$

- Can't be known but we can produce cutoffs so $E(FDP) < q$
- Benjamini and Hochberg (1995):
  - Rank the n p-values, smallest to largest, $p_1 \ldots p_n$.
  - p-value cut-off = max $(p_k : p_k \leq \frac{qk}{n})$

False Discovery Rate

$$\text{False discovery Proportion} = \frac{\#\text{ of FP}}{\#\text{ of Sig. Results}}$$

- Can't be known but we can produce cutoffs so $E(FDP) < q$
- Benjamini and Hochberg (1995):
  - Rank the n p-values, smallest to largest, $p_1 \ldots p_n$.
  - p-value cut-off = max $(p_k : p_k \leq \frac{qk}{n})$
  - All p-values below that accepted

False Discovery Rate

$$\text{False discovery Proportion} = \frac{\#\text{ of FP}}{\#\text{ of Sig. Results}}$$

- Can't be known but we can produce cutoffs so $E(FDP) < q$
- Benjamini and Hochberg (1995):
  - Rank the n p-values, smallest to largest, $p_1 \ldots p_n$.
  - p-value cut-off = max $(p_k : p_k \leq \frac{qk}{n})$
  - All p-values below that accepted
  - Caveat: Assumes independence

Other Ways of Being Wrong

- Sampling

Other Ways of Being Wrong

- Sampling
- Changing data generating process over time

Other Ways of Being Wrong

- Sampling
- Changing data generating process over time
- Confounding variables (‘data leakage’)

Other Ways of Being Wrong

- Sampling
- Changing data generating process over time
- Confounding variables (‘data leakage’)
- Coding and computational errors