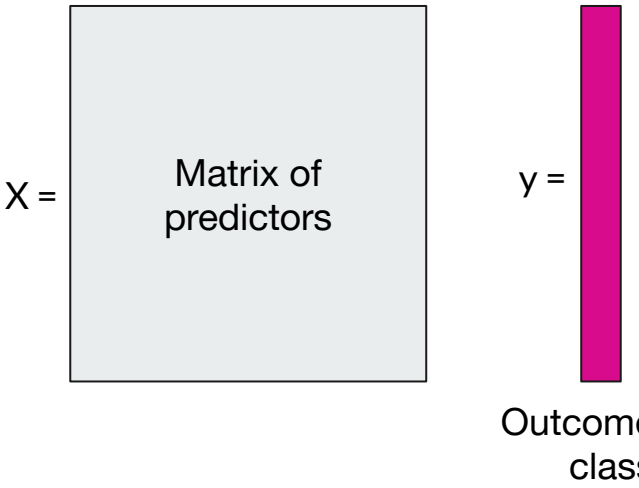
Chapter 16: Feature Selection

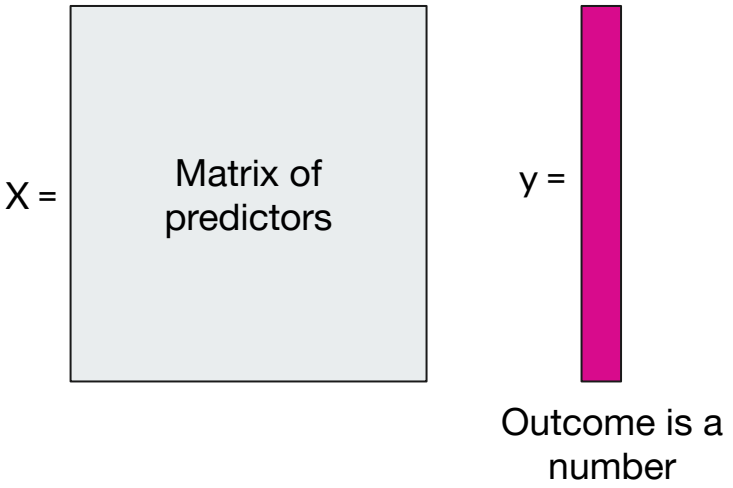
Modern Clinical Data Science
Chapter Guides
Bethany Percha, Instructor

How to Use this Guide

- Read the corresponding notes chapter first
- Try to answer the discussion questions on your own
- Listen to the chapter guide (should be 30 min, max) while following along in the notes



Classification

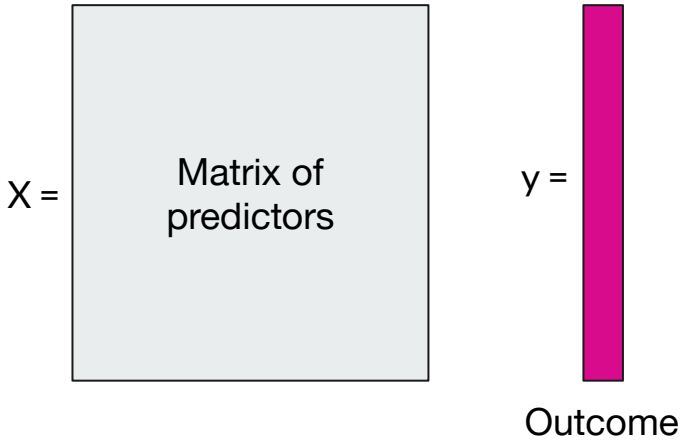


Regression

So far…

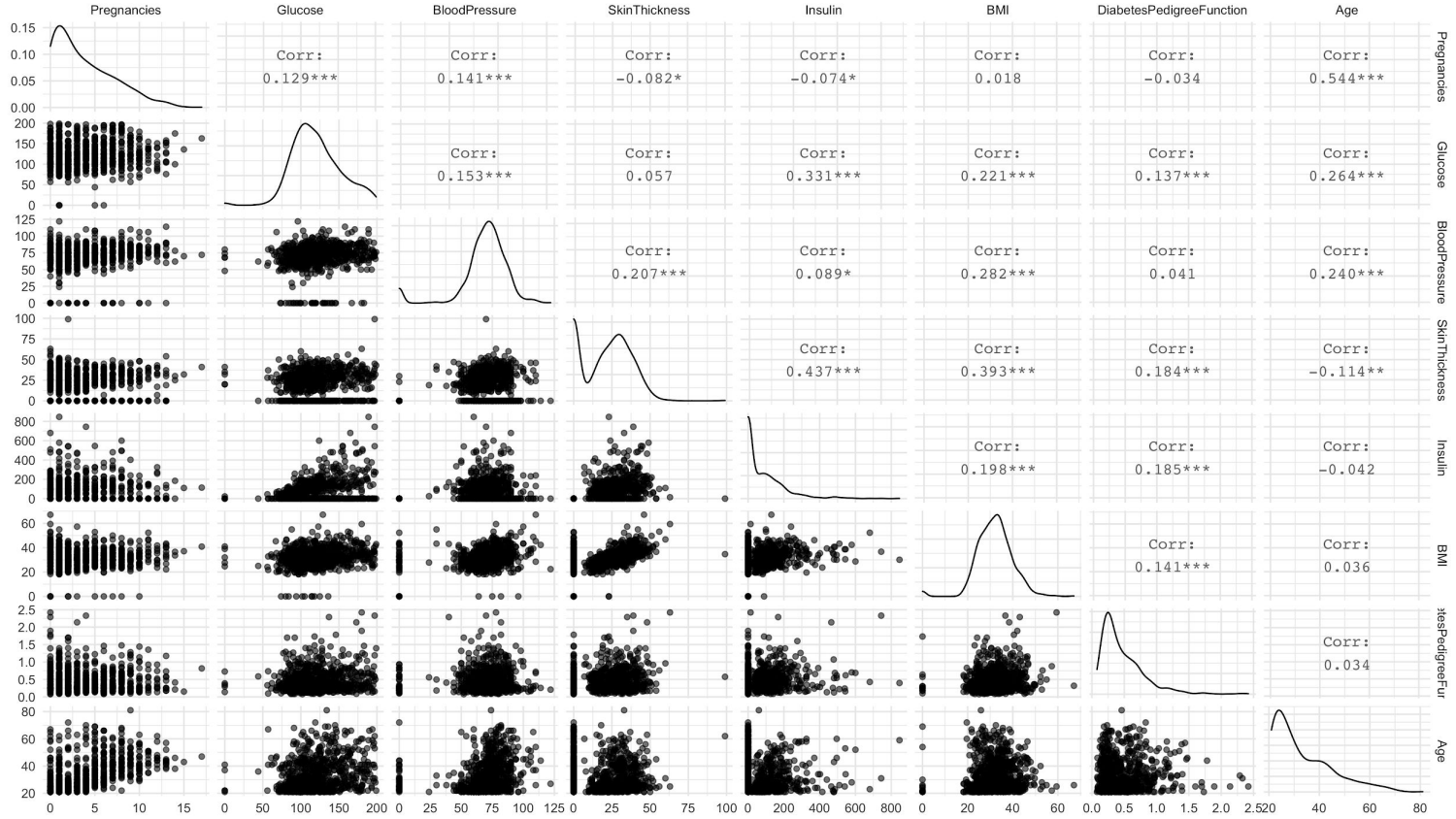
- Regression
- K-Nearest Neighbors (KNN)
- Decision trees
- Random forests
- Boosted decision trees

What happens when we have too many features, or correlated features?



Predictor	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration in a two-hour oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	Two-hour serum insulin (μU/mL)
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function (developed by research team; described in paper)
Age	Age in years

Correlograms



Univariate Analyses

Predictor	Unadjusted Coefficient	Unadjusted Odds Ratio	Unadjusted P-value
Pregnancies	0.137	1.147	<0.001
Glucose	0.038	1.039	<0.001
BloodPressure	0.007	1.007	0.073
SkinThickness	0.010	1.010	0.039
Insulin	0.002	1.002	<0.001
BMI	0.094	1.100	<0.001
DiabetesPedigreeFunction	1.083	2.953	<0.001
Age	0.042	1.043	<0.001

Predictor	Adjusted Coefficient	Adjusted Odds Ratio	Adjusted P-value
Pregnancies	0.123	1.131	<0.001
Glucose	0.035	1.036	<0.001
BloodPressure	−0.013	0.987	0.011
SkinThickness	0.001	1.001	0.929
Insulin	−0.001	0.999	0.186
BMI	0.090	1.094	<0.001
DiabetesPedigreeFunction	0.945	2.573	0.002
Age	0.015	1.015	0.111

Question 16.2

How can the odds ratio for Insulin be so close to 1.0 yet its p-value so low? (Hint: See Section 12.5.)

Question 16.3

Why might the coefficient and p-value for SkinThickness change so much in the shift from unadjusted to adjusted?

Filter Methods

Select subsets of variables as a preprocessing step, independently of the supervised learning model that will eventually be implemented.

- Any kind of univariate model (e.g. univariate logistic or linear regression)
- Any kind of hypothesis test (e.g. t-test, chi-squared test; see Chapter 6)
- Any kind of correlation coefficient (e.g. Pearson, Spearman)

- Mutual information[3]

$$MI(X_i, Y) = \sum_x \sum_y P(X_i = x, Y = y) \log \frac{P(X_i = x, Y = y)}{P(X_i = x)P(Y = y)}$$

- Variance thresholding (simply remove features with low variance)

Question 16.4

If you wanted to use the univariate logistic regression models above in Section 16.3 as a filter for a downstream model (potentially not even multivariate logistic regression - it could be a decision tree, etc.), how would you rank them and how would you decide on an appropriate cutoff?

Question 16.5

How would you apply a filter-based selection method in a case where you had dozens of different predictors of different types (e.g. some categorical, some binary, some numeric)?

Question 16.6

How might you choose the appropriate threshold for a filter-based method in a data-driven way?

Question 16.7

What is problematic about testing each potential feature, one at a time?

Wrapper Methods

Use a search algorithm to traverse the space of possible features.

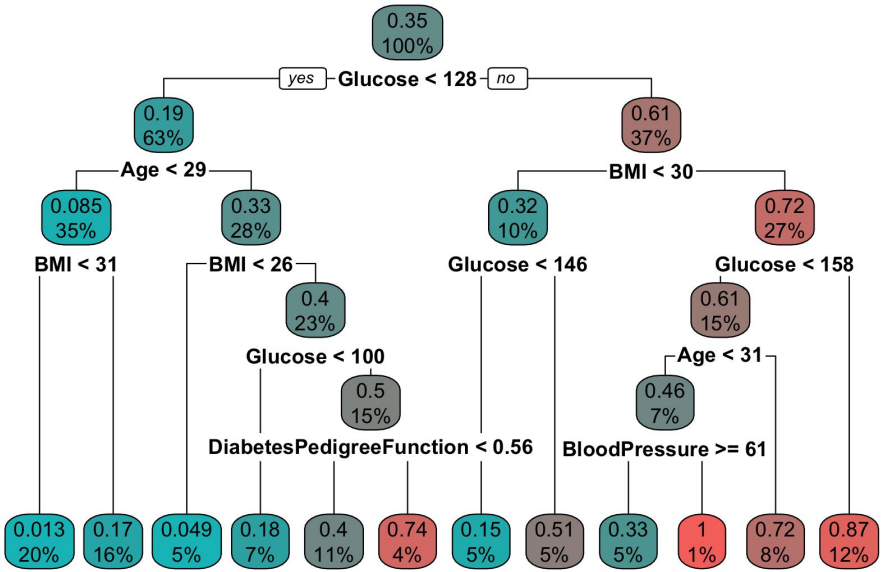
- **Exhaustive search.** Try all possible subsets of features. If there are m features, this means trying 2^m possible subsets.
- **Forward selection.** Start with a baseline (e.g., intercept only) model. Add in each of m possible predictors individually and take the best one based on some performance criterion. Repeat, adding one predictor at each step, until the performance criterion stops getting better or you run out of predictors.
- **Backward elimination.** Start with a complete model (all predictors included). Try removing each predictor and take the one whose removal causes the performance criterion to increase the most. Repeat, removing one predictor at each step, until the performance criterion stops getting better or you are left with no predictors (null model).
- **Forward-backward selection.** A combination of forward selection and backward elimination.
- **Simulated annealing.** Add or remove predictors with some probability depending on how well the model is doing. At each stage, if the new model is better, accept it; it becomes the new baseline. If the new model is worse, accept it with some probability, p, that decreases over time according to a "cooling schedule". This helps prevent the variable selection process from getting stuck in local optima.

Question 16.8

Why is exhaustive search problematic for almost any reasonably sized m?

Note: Forward and backward stepwise selection for the Pima dataset are shown in the chapter.

Embedded Methods



Another example: regularization.

See Chapter 17!