Classification

STA 198: Introduction to Health Data Science

Yue Jiang

July 17, 2020

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

Logistic regression

Last time, we talked about how we can use logistic regression to describe relationships between predictors and the odds of events occurring.

This was an example of binary regression, where our outcome was a categorical variable.

Logistic regression

By working backwards from the predicted log-odds, we can obtain *predicted probabilities* of “success” for a binary variable. Thus, we have the predicted probability of being a “success.”

By instituting a cut-off value (e.g., if the probability is greater than 0.5), we can create a classifier.

This can be extended to more than 2 categories, but this is beyond the scope of our course (for the curious: e.g., multinomial regression).

Strengths of logistic regression

- Linear model of transformation of binary response
- Straightforward interpretation of coefficients
- Odds ratios have intuitive appeal in many health/biomedical fields
- Ability to handle both continuous and categorical predictors
- Can quantify degree of uncertainty around prediction
- Many extensions

Drawbacks of logistic regression

- "Decision boundary" between classes is constrained to be linear (more on this later)
- Requires additional assumptions regarding independence and the transformation used
- Coefficient estimates may be unreliable if predictors are highly correlated (colinearity)

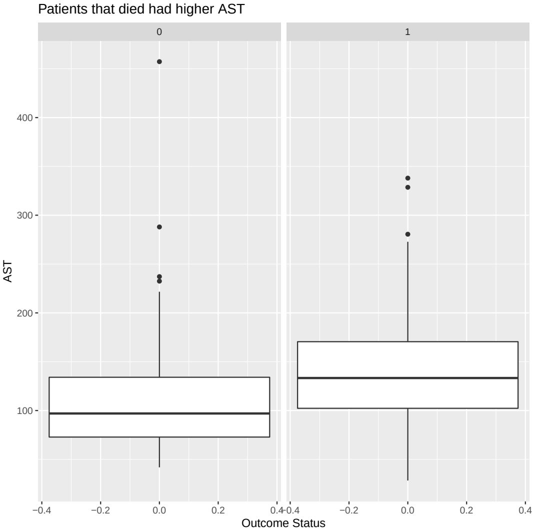
k-nearest neighbors

Recall: PBC data from the Mayo Clinic. For now, goal is to classify whether a patient dies at the end of follow-up, based on a new set of predictors:

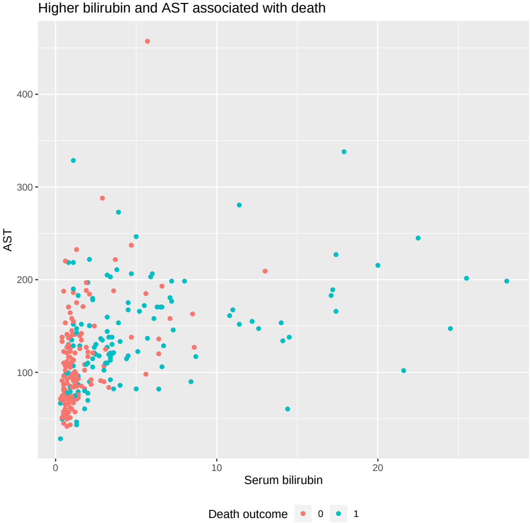
- Age
- AST (aspartate aminotransferase) levels
- Serum bilirubin levels
- Platelet count
- Standardized blood clotting time

Note that we are treating these as *continuous* predictors. k-nearest neighbors relies on the notion that "similar observations are similar"

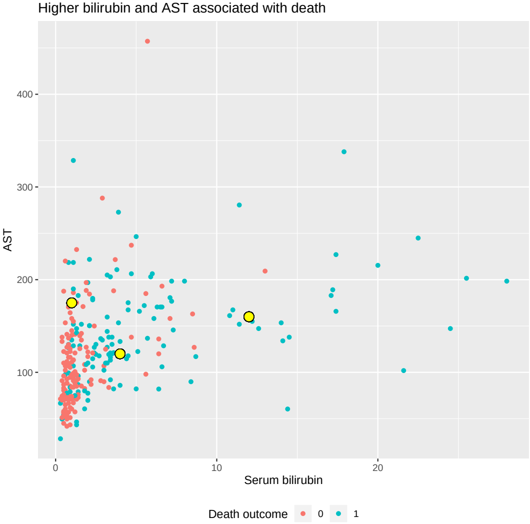
Visualizing AST and death



Visualizing AST, bilirubin, and death



Some hypothetical patients...



The general idea

Given a new data point, predict its class status by taking the plurality vote of its k nearest neighbors in terms of *their* class memberships.

Distance functions

It's straightforward enough to visualize the "nearest" neighbors if we're only using two predictors. But suppose we want to use all five predictors of interest. How can we do that?

Suppose $\mathbf{x}$ and $\mathbf{y}$ are n-dimensional vectors. Then the Euclidean distance between them is given by

$$\begin{aligned} D(\mathbf{x}, \mathbf{y}) &= \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \end{aligned}$$

Other choices of distance function are available.

How do we choose k?

- Trade-off between reducing variance, computational expense, and sharpness of class boundary
- Larger k results in "fuzzier" class boundaries; smaller k results in sharper class boundaries
- A commonly used approach is to use the square root of the sample size
- Using cross-validation to choose k is often a good idea in practice, but we will not cover that in this course

Classifying a hypothetical person

Suppose we have a hypothetical patient with the following features. Would we predict them to die within the follow-up time period?

- Age: 40
- AST: 96
- Bilirubin: 1.6
- Platelet: 266
- Prothrombin time: 10.4

We'll calculate their *k* nearest neighbors in 5-dimensional Euclidean space, and take the plurality vote as whether we predict they will "succeed" (die).

Strengths of k-NN

- Intuitive to understand and implement
- Decision boundary can have an arbitrary shape
- Virtually assumption-free
- Easy to extend to multi-class problems (keep on taking the plurality vote)
- Can be extended to add flexibility (e.g., weighting votes based on distance)

Drawbacks of k-NN

- Unbalanced class sizes are difficult to resolve, since rare classes will be dominated by sheer number of other observations
- Computationally intensive
- Sensitive to high variance predictors, irrelevant predictors, and outliers
- Completely ignores “far away” points
- Requires that predictors can be compared on the same scale
- Must determine k and distance function (how can we calculate a distance for categorical predictors?)
- Cannot deal with missing values