VANDERBILT
PEABODY COLLEGE

The Problem of Classification

Will Doyle

Classification

- Predicting group membership
- Simplest version: binary dependent variable
- Can also be used for ordinal data (ranked, but nonnumeric)
- Can also be used for categorical data (any of a large number of discrete groups)

Conditional Mean as a Classifier

- We can continue to use conditional means as a classifier.
- It gives surprisingly good answers.
- But the curse of dimensionality is always with us.

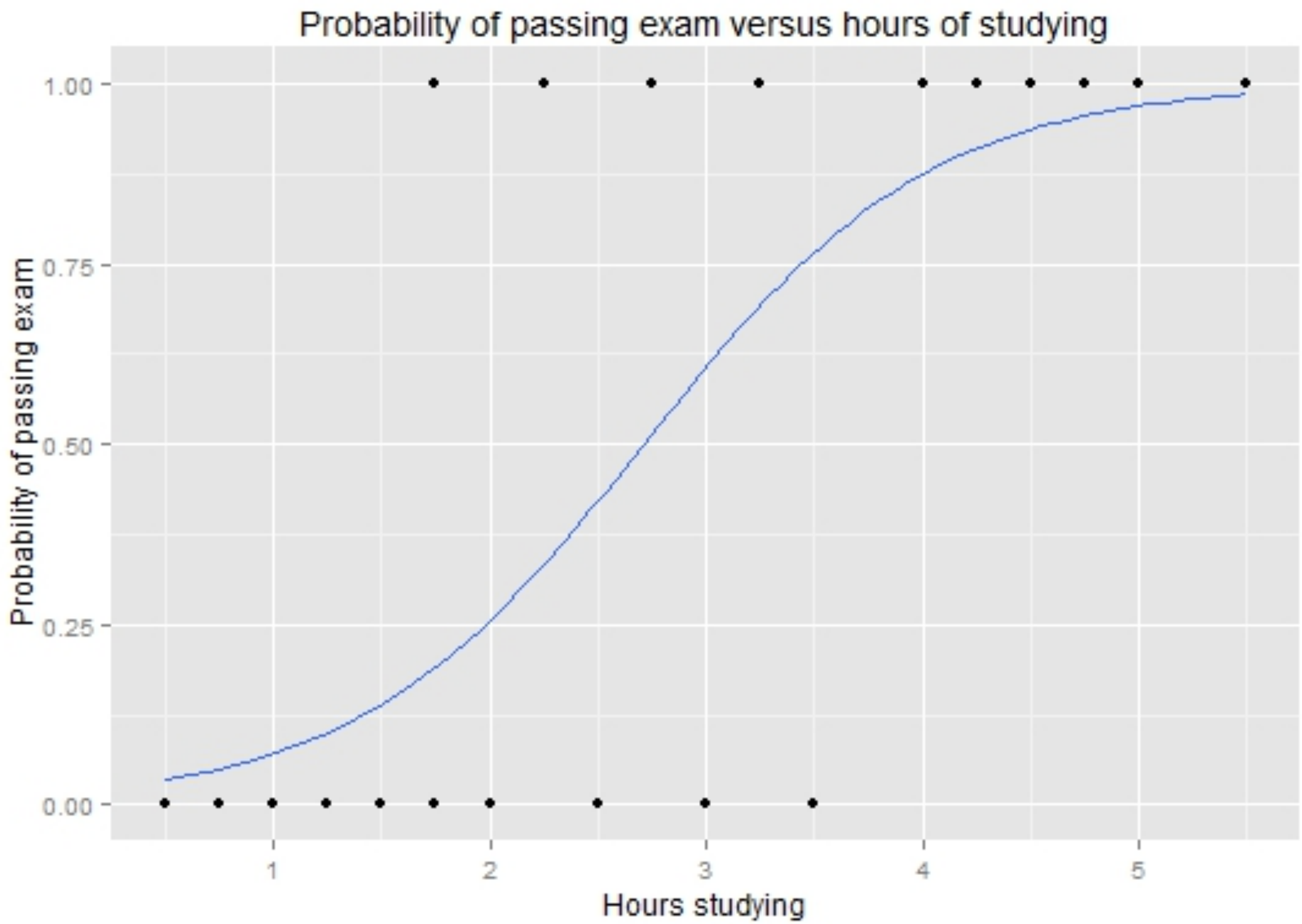
Linear Regression as a Classifier

- Linear regression can be used as a classifier for binary group membership.
- It is known as a “linear probability model.”
- Real-world performance is often quite good.
- It can generate probabilities lower than 0, greater than 1 (not good).
- Performance is poor when there are a lot of respondents in one group or another ($\Pr(y = 1) < .05$, $\Pr(y = 1) > .95$)

Logistic Regression as a Classifier

- One of a family of generalized linear models
- Uses maximum likelihood estimation (doesn’t directly compute estimates, but chooses estimates which maximize probability of data)
- Parameter estimates can’t be directly interpreted on a probability scale

Visualizing Logistic Curve



Source: https://upload.wikimedia.org/wikipedia/commons/6/6d/Exam_pass_logistic_curve.jpeg

VANDERBILT
PEABODY COLLEGE

Evaluating Classifiers

Sensitivity and Specificity

Will Doyle

Evaluating Classifiers

- A good classifier will accurately predict group membership.
- There are well-known trade-offs in creating classifiers.
- Some classifiers are very good at predicting group membership. This is called having high sensitivity.
- Some classifiers are very good predicting when someone is **not** in a group. This is called having high specificity.

Example: Purchasing a Product

	Actual outcome: Purchased	Actual outcome: Did not purchase
	25	75

Example: Purchasing a Product Classifier Is Sensitive but Not Specific

	Actual outcome: Purchased	**Actual outcome: Did not purchase**
Predicted outcome: Purchased	25	75
Predicted outcome: Did not purchase	0	0

Example: Purchasing a Product Classifier Is Specific but Not Sensitive

	Actual outcome: Purchased	**Actual outcome: Did not purchase**
Predicted outcome: Purchased	0	0
Predicted outcome: Did not purchase	25	75

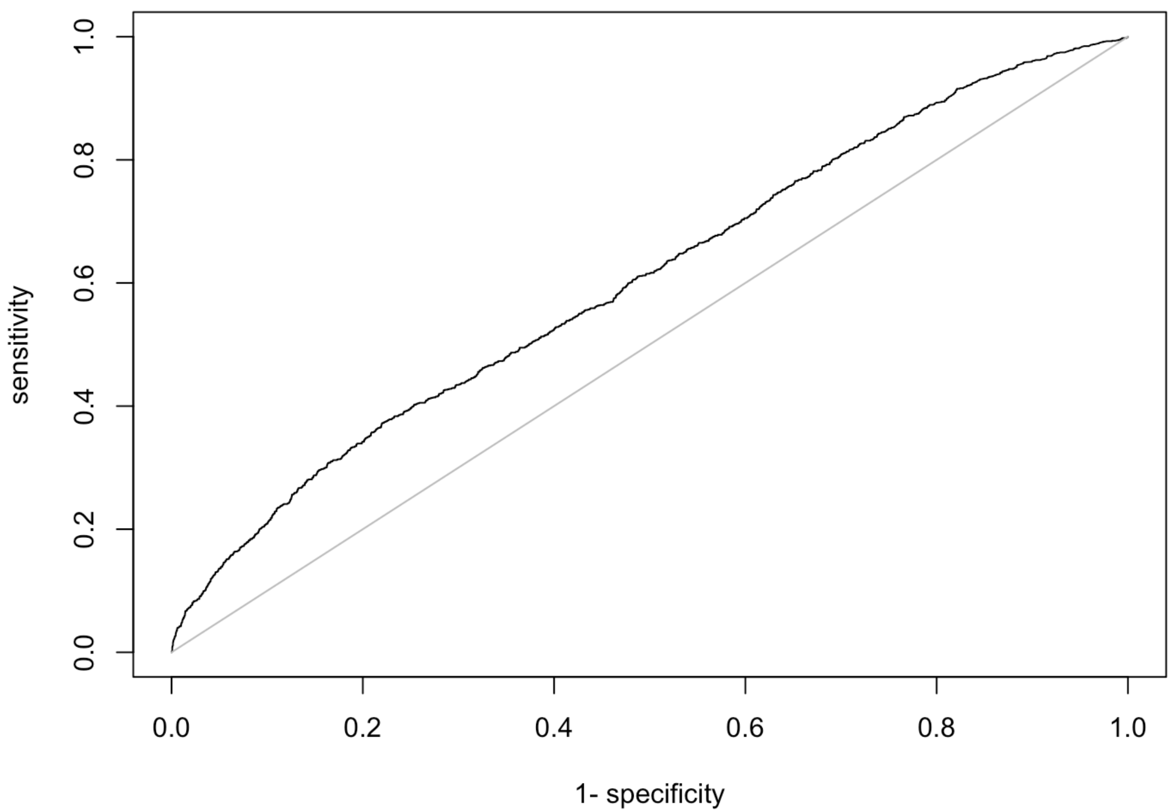
Sensitivity and Specificity

- A good classifier will be both sensitive **and** specific.
- But it’s not easy.
- We can vary the model of course.
- We can also vary the classification threshold from 0 to 1.

Receiver Operator Characteristic Curve

- Calculates sensitivity **and** 1—specificity over a range of classification thresholds.
- Thresholds go from 0–1.
- The measure of AUC (area under curve) measures how well the model classifies at every threshold.
- A perfect classifier will have an AUC of 1 (never actually happens).
- A random classifier will have an AUC of .5.

ROC and AUC



VANDERBILT
PEABODY COLLEGE