

Biostat 202: Opportunities and Challenges of “Big Data”.

Machine Learning 2: Supervised learning [classification]

Aaron Wolfe Scheffler

Division of Biostatistics

Department of Epidemiology and Biostatistics

Announcements



Assignment 2 due
today

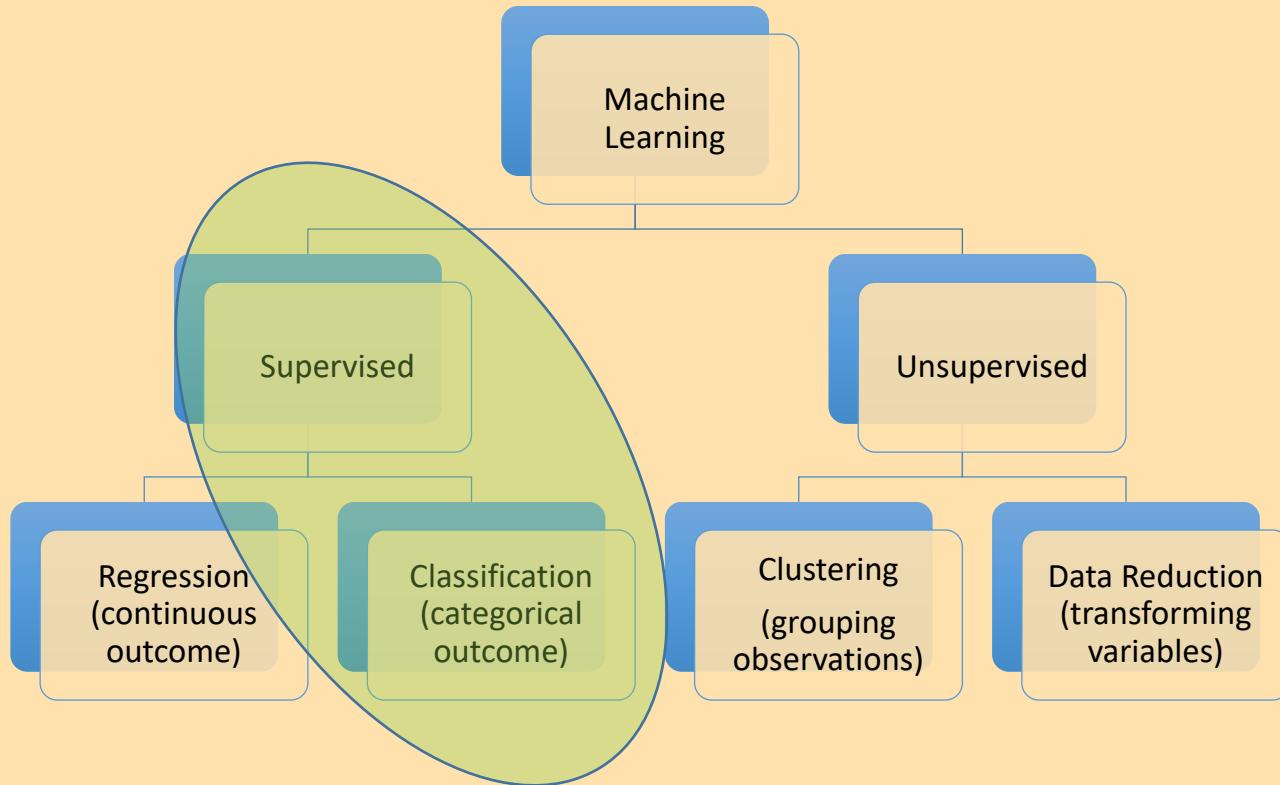


Project Component
1 due Sunday, 8/16
@ midnight

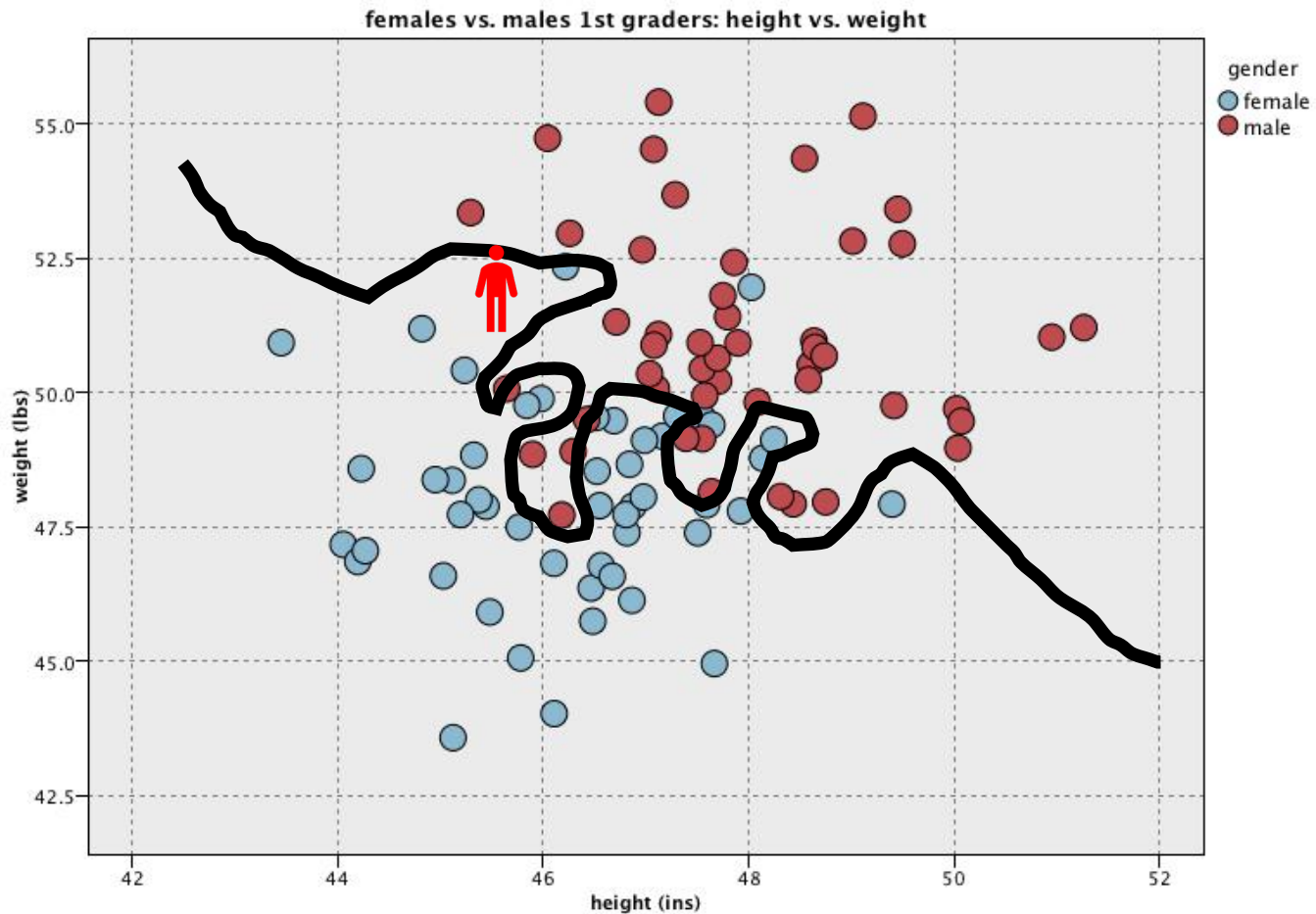
Overview of this lecture

- Introduction to classification
- Evaluation metrics for categorical data
- Validation and testing, continued
- Logistic regression
- Heart disease case study

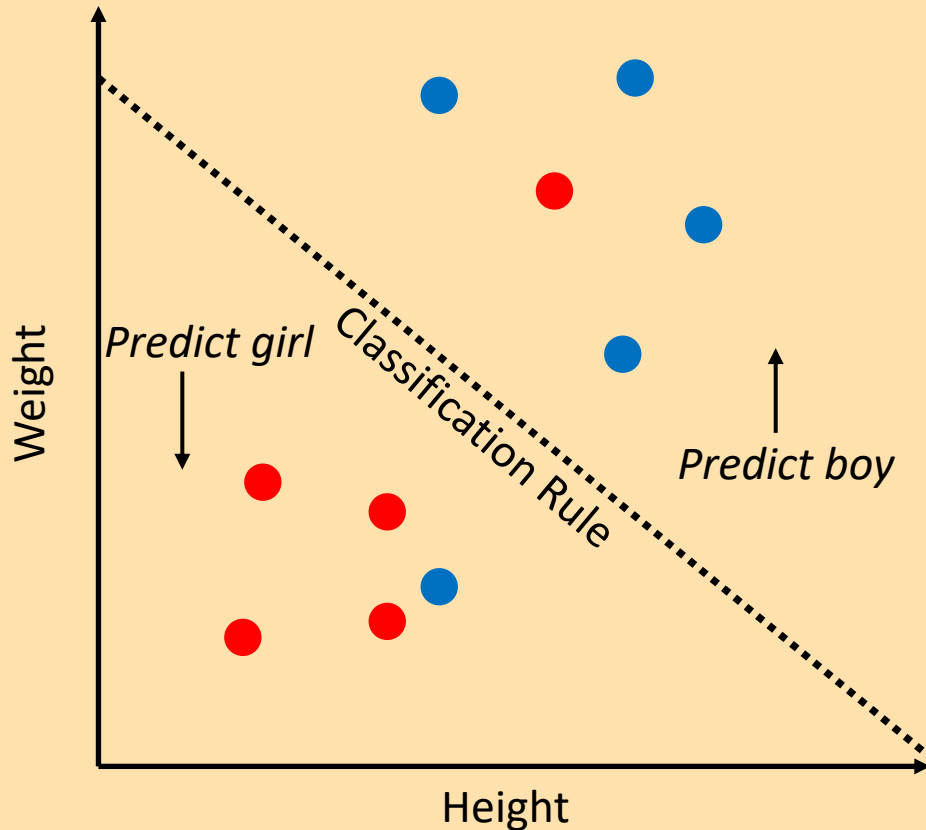
Classification



2D classification example



Definitions



- True Positive (TP) – observation is positive and was predicted as positive
- True Negative (TN) – observation is negative and was predicted as negative
- False Positive (FP) – observation is negative and was predicted as positive
- False Negative (FN) – observation is positive and was predicted as negative
- Classification accuracy – proportion of observations classified correctly

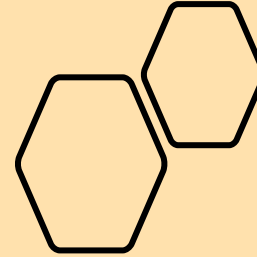
$$= (\#TP + \#TN)/N$$

$$= (4 + 4)/10 = 80\%$$

Evaluation metrics for categorical data

- Possible evaluation metrics:
 - *Classification Accuracy* – or conversely, the *classification error rate* = $1 - \text{classification accuracy}$
 - *Cost-Benefit* – some errors (e.g. FP, FN) are counted more than other, often context dependent
 - Sensitivity, specificity, precision
 - Area Under Receiver Operating Characteristic (ROC) Curve (AUC)
- Guiding questions in selecting evaluation metrics
 - What is my analytical goal, i.e. research question?
 - What metric can best help me evaluate how reliable by predicted results are in the context of my analytical goal?

Evaluation Metric Practice!



Download “Biostat
202 Evaluation
Metric Exercise”
document from CLE

Logistic Regression

Logistic regression

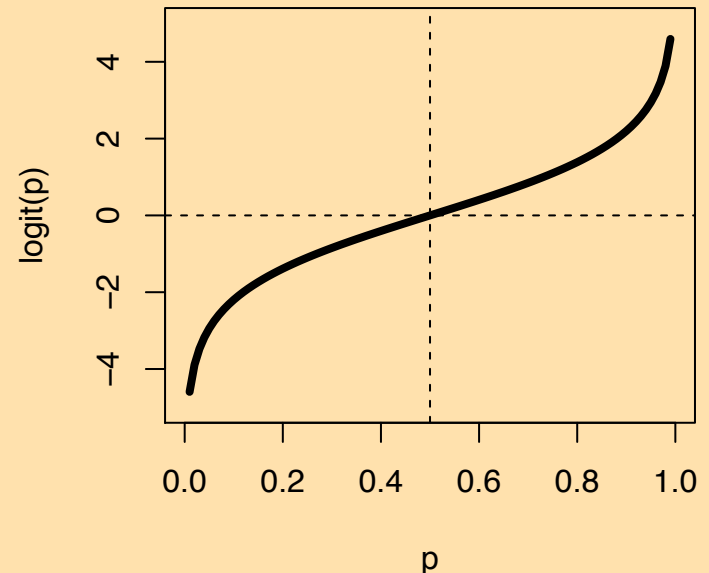
p must be between 0 and 1 (probability)

We can instead take the *odds* which is between 0 and infinity: $\frac{p}{1-p} > 0$ but still not quite what we need...

to get values in the range 0 to 1, we take the log of the odds (*logit function*) and get the logistic regression equation:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = a * \text{height} + c$$

We can extend logistic regression to multiple predictors just as we did for linear regression.



Logistic regression for classification

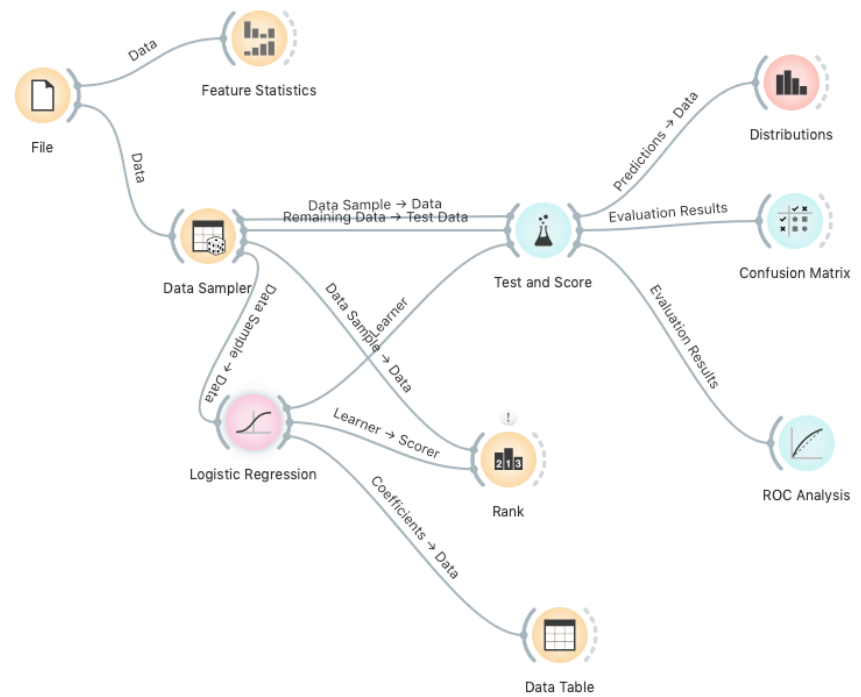
- When fitting logistic regression we get a model that relates predictors to a class probability
- To convert the probability to a class we need to threshold the probability
- E.g. if probability $p \leq 0.5$ then class as girl, otherwise if $p > 0.5$ class as boy

Ongoing case study – Heart data

- Data:
 - $n = 303$
 - 14 attributes
 - **Target:**
 - **AHD: diagnosis of heart disease (1 = yes, 0 = no)**
- Predictors:
- Age: age (years)
 - Sex: gender
 - ChestPain: chest pain type
 - Chol: cholesterol
 - Fbs: fasting blood sugar >120 mg/dL
 - restecg: resting electrocardiographic results
 - MaxHR: maximum heart rate achieved
 - ExAng: exercise induced angina
 - Oldpeak: ST depression induced by exercise
 - Slope: slope of peak exercise ST segment
 - Ca: # of major vessels colored by flouroscopy
 - Thal: Thallium stress test
 - SysBp: systolic blood pressure

Goal: Predict (probability) of heart disease using logistic regression.

Logistic Regression Orange Demo



QUESTIONS?

A thin, vertical white line is positioned to the right of the word "QUESTIONS?". It extends from approximately the middle of the text's vertical height down to about one-third of the way down the slide.

Next Monday – more classification!

- K-nearest neighbors
 - Parameter tuning
 - Training-validation-test
 - Cross-validation
- Classification trees (recursive partitioning)
- Support vector machines
- Artificial neural networks
- Comparison of classifiers

LAB



Download
Assignment 3 from
CLE



Download
Framingham.csv
from CLE