Methods & Characteristics

The three methods:

- Naïve rule
- Naïve Bayes
- K-nearest-neighbor

Common characteristics:

- Data-driven, not model-driven
- Make no assumptions about the data

Naïve Rule

- Classify all records as the majority class
- Not a “real” method
- Introduced so it will serve as a benchmark against which to measure other results

Naïve Bayes

Naïve Bayes: The Basic Idea

- For a given new record to be classified, find other records like it (i.e., same values for the predictors)
- What is the prevalent class among those records?
- Assign that class to your new record

Usage

- Requires categorical variables
- Numerical variable must be binned and converted to categorical
- Can be used with very large data sets
- Example: Spell check – computer attempts to assign your misspelled word to an established “class” (i.e., correctly spelled word)

Exact Bayes Classifier

- Relies on finding other records that share same predictor values as record-to-be-classified.

- Want to find “probability of belonging to class *C*, given specified values of predictors.”

- Even with large data sets, may be hard to find other records that **exactly match** your record, in terms of predictor values.

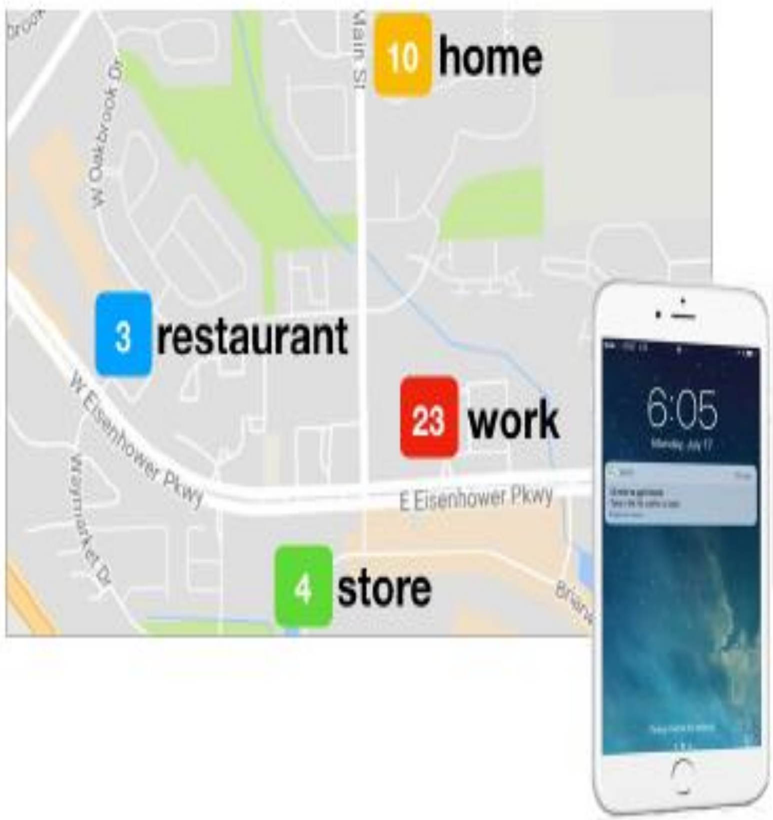
Solution – Naïve Bayes

- Assume independence of predictor variables (within each class)

- Use multiplication rule

- Find same probability that record belongs to class C, given predictor values, <u>without</u> limiting calculation to records that share all those same values

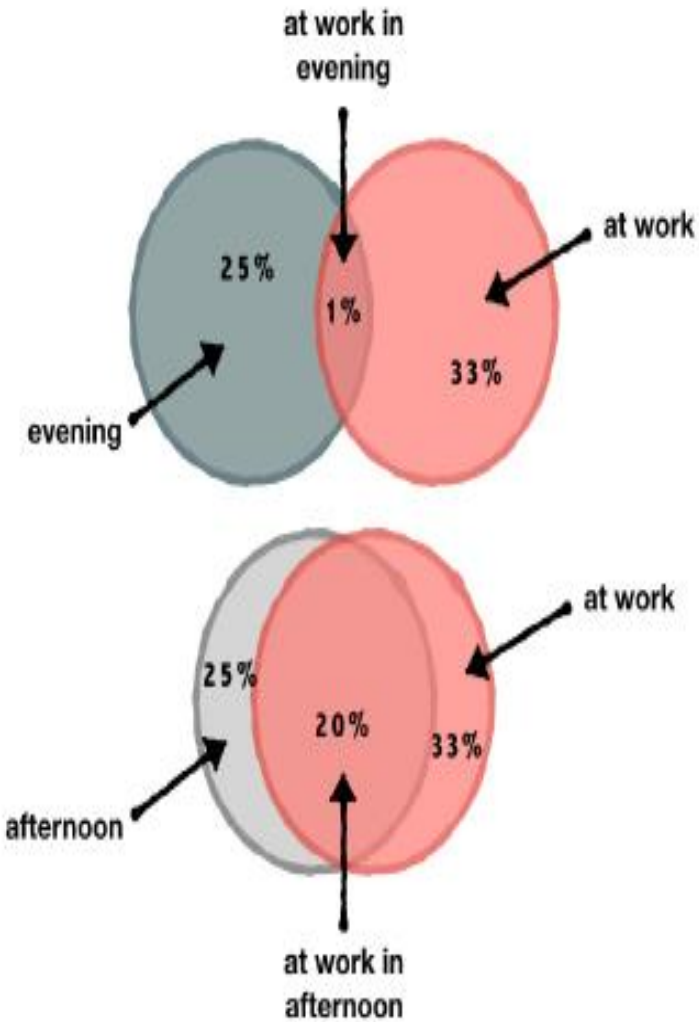
Estimating probability



The **probability** of A is denoted P(A)

- P(work) = 23 / 40 = 57.5%
- P(store) = 4 / 40 = 10.0%

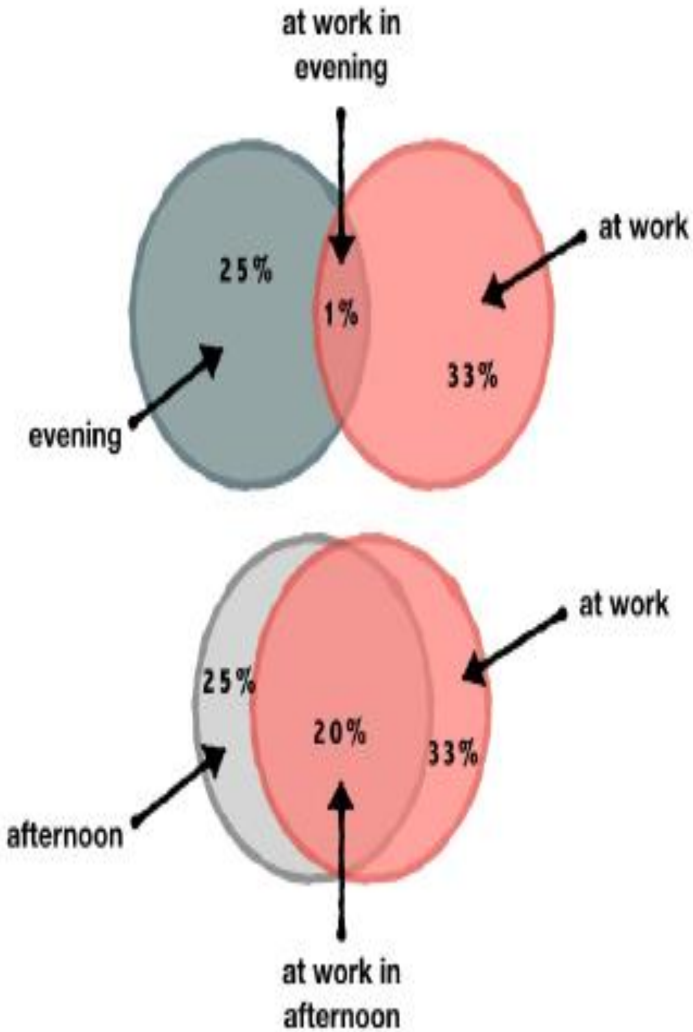
Joint probability and independent events



The **joint probability** of events A and B is denoted P(A and B)

- P(work and evening) = 1%
- P(work and afternoon) = 20%

Conditional probability and dependent events



The **conditional probability** of events A and B is denoted P(A | B)

- P(A | B) = P(A and B) / P(B)
- P(work | evening) = 1 / 25 = 4%
- P(work | afternoon) = 20 / 25 = 80%

Example: Financial Fraud

Target variable: Audit finds fraud, no fraud

Predictors:

- Prior pending legal charges (yes/no)
- Size of firm (small/large)

Charges?	Size	Outcome
y	small	truthful
n	small	truthful
n	large	truthful
n	large	truthful
n	small	truthful
n	small	truthful
y	small	fraud
y	large	fraud
n	large	fraud
y	large	fraud

Exact Bayes Calculations

- Goal: classify (as "fraudulent" or as "truthful") a small firm with charges filed
- There are 2 firms like that, one fraudulent and the other truthful
- P(fraud|charges=y, size=small) = ½ = 0.50
- Note: calculation is limited to the two firms matching those characteristics

Naïve Bayes Calculations

- Goal: Still classifying a small firm with charges filed
- Remember we are trying to model

$$\pi_j p_j(x)$$

- Assuming independence of the features in each class we write

$$\pi_j p_j(x_1,..x_p) = \pi_j p_j(x_1) \times ..... \times p_j(x_p)$$

Naïve Bayes Calculations

In the present example, compute these quantities:

- Proportion of "charges = y" among frauds, times proportion of "small" among <u>frauds</u>, times proportion frauds = 3/4 * 1/4 * 4/10 = 0.075
- Prop "charges = y" among truthfuls, times prop. "small" among <u>truthfuls</u>, times prop. truthfuls = 1/6 * 4/6 * 6/10 = 0.067

P(fraud|charges, small) = 0.075/(0.075+0.067)

= 0.53

Making predictions with Naive Bayes

```r
# building a Naive Bayes model
library(naivebayes)
m <- naive_bayes(location ~ time_of_day, data = location_history)
```

```r
# making predictions with Naive Bayes
future_location <- predict(m, future_conditions)
```

Naïve Bayes, cont.

- Note that probability **estimate** does not differ greatly from **exact**
- All records are used in calculations, not just those matching predictor values
- This makes calculations practical in most circumstances
- Relies on assumption of independence between predictor variables within each class

Independence Assumption

- Not strictly justified (variables often correlated with one another)
- Often “good enough”

Advantages

- Handles purely categorical data well
- Works well with very large data sets
- Simple & computationally efficient

Shortcomings

- Requires large number of records
- Problematic when a predictor category is not present in training data
  - Assigns 0 (zero) probability of response, ignoring information in other variables

On the other hand…

- Probability rankings are more accurate than the actual probability estimates
  - Good for applications using lift (e.g. response to mailing), less so for applications requiring probabilities (e.g. credit scoring)