Naïve Bayes

Naïve Bayes derivation

Feature vector

- On $\boldsymbol{x}$ predict $\text{argmax}_y\ P(y \mid \boldsymbol{x})$

$$= \text{argmax}_y\ P(\boldsymbol{x} \mid y)\ P(y) / P(\boldsymbol{x})$$

$$= \text{argmax}_y\ P(\boldsymbol{x} \mid y)\ P(y)$$

- Naïve independence assumption:

$$P(\boldsymbol{x} \mid y) = \prod_j P(x_j \mid y)$$

Attributes conditionally
Independent given y

- Predict the label y maximizing

$$P(y) \prod_j P(x_j \mid y)$$

these distributions are the model

- Uses generative model: pick y then generate $\boldsymbol{x}$ based on y

Need data to estimate:

- P(y) distribution

- For each class y, for each feature x_j
  need P(x_j | y) distributions

- All these distributions “1-dimensional”

Naïve Bayes example using max likelihood estimates (empirical counts)

- Data: (boolean)

$\boldsymbol{x}$	y
T,T	+1
T,F	+1
F,T	+1
F,T	+1
F,F	-1
T,F	-1
F,T	-1

- Predict on $\boldsymbol{x}$=(T,F) using max likelihood estimates from data

P(y = +1) = 4/7; P(y = -1) = 3/7

P(x_1=T | y=+1) = 1/2

P(x_2=F | y=+1) = 1/4

P(x_1=T | y=-1) = 1/3

P(x_2=F | y=-1) = 2/3

For "+1": (4/7)(1/2)(1/4) = 1/14

For "-1": (3/7)(1/3)(2/3) = 2/21

Predict "-1"

Naïve Bayes example using max likelihood estimates

- Data: (boolean)

x	*y*
T,T	+1
T,F	+1
F,T	+1
F,T	+1
F,F	-1
T,T	-1
F,F	-1

- Predict on ***x***=(T,F) using max likelihood estimates from data

$P(y = +1) = 4/7$; $P(t = -1) = 3/7$

$P(x_1=T \mid y=+1) = 1/2$

$P(x_2=F \mid y=+1) = 1/4$

$P(x_1=T \mid y=-1) = 1/3$

$P(x_2=F \mid y=-1) = 2/3$

For "+1": $(4/7)(1/2)(1/4) = 1/14$

For "-1": $(3/7)(1/3)(2/3) = 2/21$

Predict "-1", even on +1 example!

Naïve Bayes discussion

- Straight from data, no searching
  - But need to estimate class conditional prob's – the probabilities of feature-values given the class
- Successful applications include:
  - Medical diagnosis
  - Classifying text (Joachims, 1996) 89% accuracy for identifying source from 20 newsgroups (1000 documents each group, 2/3 train 1/3 test)
  - Newsweeder (Lang, 1995) interesting articles up from 16% to 59% after filtering

Naïve Bayes Issues

1. Conditional independence optimistic, but…

   Don’t have to get probabilities right, just the predictions – also decision threshold tuning
2. What if an attribute-value pair not in training set for all labels?
   - Use Laplace smoothing
3. Numeric Features: use Gaussian or other density (Poisson, exponential) (degeneracy issue?)
4. Attributes for text classification?
   - Bag of words model

Naïve Bayes for Text

(see Mitchell's book)

- Let V be the vocabulary (all words/ symbols in all training documents)
- For each class y, let Docs_y be the concatenation of all docs labeled y
- For each word w in V, let $\#w(\text{Docs}_y)$ be # of times w occurs in Docs_y
- Set $P(w \mid y)$ to:

$$(\#w(\text{docs}_y) + 1) \ / \ (|V| + \sum_w \#w(\text{docs}_y))$$

Laplacian smoothing

Naïve bayes for text (2)

- Predict on new document $\boldsymbol{x}$ with class y maximizing

$$P(y) \prod_{\mathrm{w\ in\ } \boldsymbol{x}} P(\mathrm{w} \mid y)$$

Note: repeated words multiplied in multiple times (multinomial model)

Feature vector $\boldsymbol{x}$ is vector of counts

Exercise:

- Repeat slide 3 example using Laplacian probability estimates. Calculate the “vote” for each of the two classes for the new instance ***x***=(T,F).
- Use Naïve Bayes in Weka for iris2.arff (iris.arff)

- Data: (boolean)

T,T	+1
T,F	+1
F,T	+1
F,T	+1
F,F	-1
T,F	-1
F,T	-1