Naïve Bayes

Adrian Barbu

Classification Using Bayes Rule

- Training set

$\overbrace{\text{Sky Temp Humid Wind Water Forecst}}^{X}$ $\overbrace{\text{EnjoySpt}}^{Y}$

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

- Direct (discriminative) approach: Learn $P(Y|X)$
  - Might not have enough training data
  - Might just want to try something else
- Bayesian approach: Use Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Learn $P(X|Y), P(Y)$

$$P(X|Y) = P(X_1, ..., X_M|Y)$$

Bayes Rule

$$P(Y|X) = \frac{P(Y,X)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

- Means

$$P(Y = y_j|X = x_i) = \frac{P(X = x_i|Y = y_j)P(Y = y_j)}{P(X = x_i)} \forall i, j$$

- Also

$$P(Y = y_j|X = x_i) = \frac{P(X = x_i|Y = y_j)P(Y = y_j)}{\sum_k P(X = x_i|Y = y_k)P(Y = y_k)} \forall i, j$$

Discriminative vs. Generative Classifiers

Learning task: learn f:X→Y or P(Y|X)

- Generative approach:
  - Assume some functional form for P(X|Y) and P(Y)
  - Learn the parameters of P(X|Y) and P(Y) from the training data
  - Generative because you *explain* the input X
  - Use Bayes rule to get P(Y|X)
  - E.g. Naïve Bayes
- Discriminative approach
  - Assuming a functional form for P(Y|X) directly,
  - Learn the parameters of P(Y|X) from the training data
  - E.g. Decision Trees, Random Forest, Regression, SVM, Boosting

Naïve Bayes

- Assume all are X_i conditionally independent given Y

$$P(X|Y) = P(X_1, ..., X_M|Y) = \prod_i P(X_i|Y)$$

- Conditional independence of X and Z given Y

$$P(X|Y, Z) = P(X|Y)$$

- E.g.

$$P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$$

- If X_1,X_2 are conditionally independent given Y

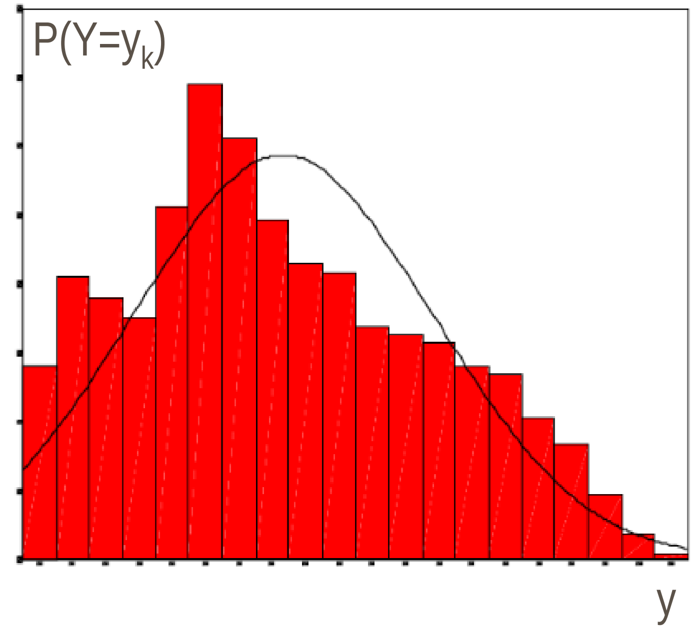
$$P(X_1, X_2|Y) = P(X_1|Y, X_2)P(X_2|Y) = P(X_1|Y)P(X_2|Y)$$

Naïve Bayes Algorithm

Training:

- For each y_k estimate $P(Y = y_k)$
  - Histogram
  - Fit a model, e.g. gaussian
- For each i and k estimate

$$P(X_i = x_{ij} | Y = y_k)$$



Classification:

- Given x^{new} find Y

$$Y = \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i = x_i^{new} | Y = y_k)$$

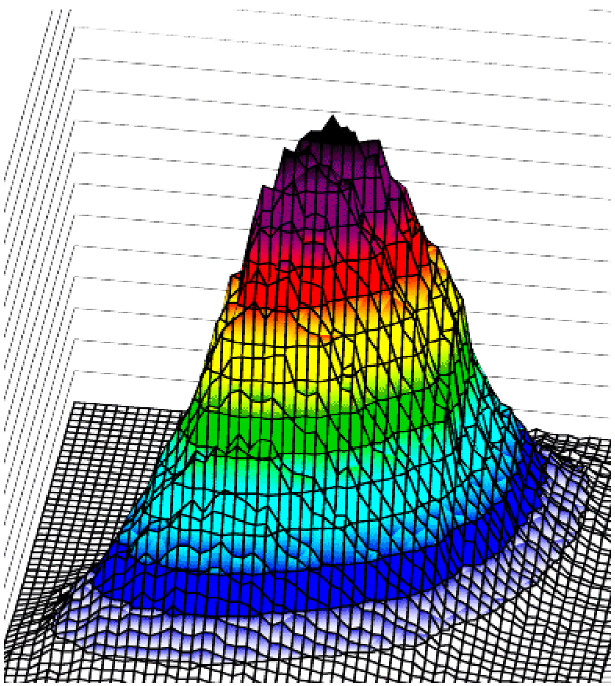
Example

- Is somebody a student (Y=1) or not (Y=-1) when we know
  - Age A in years
  - License plate style L: regular, FSU, FAMU, other
- Use naïve Bayes. X=(A,L). Must learn:
  - P(Student): 2 bin histogram
  - P(Age|Student): histogram with most values around 20
  - P(Age|Not Student): more spread out
  - P(License Plate |Student)
  - P(License Plate | Not Student)
- What if Age and License Plate are not independent conditional on student?

Removing Independence Assumption

- What if Age and License Plate are not independent conditional on student?
  - Learn two 2D histograms P(A,L|Student) and P(A,L|Non-student).
  - Requires more memory
  - Easy to overfit
  - Impractical for more than 3-4D

- Other ways
  - Parametric models



Problems with Histograms

- What if $P(X_{15}=x_0|Y=y_0)$ is zero?
  - Then the whole $P(Y=y_0|X)$ is 0
  - $Y=y_0$ could be the desired output
  - Implies overfitting

- Solution:
  - Initialize all histogram bins with 1 not 0
  - Then $P(X_{15}=x_0|Y=y_0)$ will never be 0

Text Classification

- Classify emails: Spam/ Not Spam
- Classify news:
  - Scientific
  - Business
  - Health
  - International
  - …
- Classify documents by 20 newsgroups:
  - Misc.forsale
  - Rec.auto
  - Comp.graphics
  - Sci.space
  - …

Bag of Words Approach

- Each text is a bag of words
  - Number of occurrences of each word
  - Position in text doesn't matter

MANCHESTER, New Hampshire (CNN) -- With the New Hampshire primary fast approaching, it's dead even in the race for the Democratic presidential nomination.

Hillary Clinton is tied at the top with Barack Obama.

1 of 2

Sens. Hillary Clinton of New York and Barack Obama of Illinois are tied, with each grabbing the support of 33 percent of likely Democratic primary voters in the Granite State, according to a new CNN/WMUR New Hampshire presidential primary poll conducted by the University of New Hampshire.

Former Sen. John Edwards of North Carolina is in third place with 20 percent, according to the poll, which was released Saturday afternoon, three days before the primary.

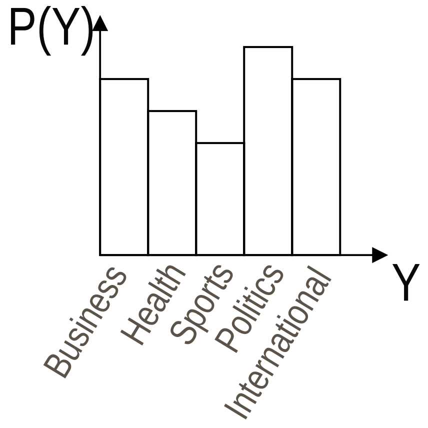
"Both **Obama** and **Edwards** appear to have benefited from the Iowa caucuses. Each picked up three points in New Hampshire. **Clinton** lost one point, since our last poll taken before the caucuses," said CNN senior political analyst Bill Schneider.

On the Republican side, John McCain has emerged the leader of the GOP pack in New Hampshire.

- about 2
- alien 0
- all 2
- Africa 0
- apple 0
- anxious 0
- ...
- Clinton 5
- ...
- republican 3
- …
- Zaire 0

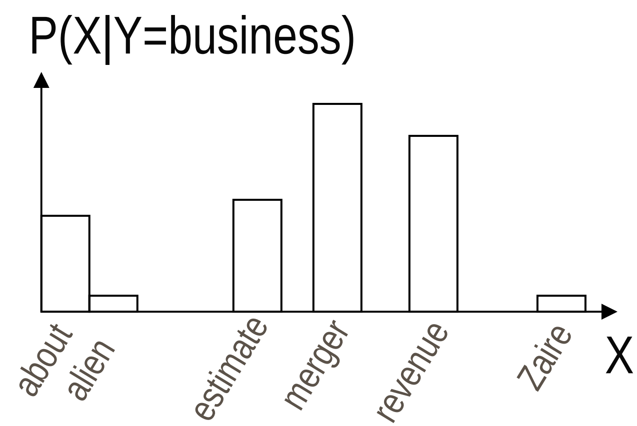
Naïve Bayes for Text Classification

- Find the *topics*:
  - Find the target labels y_i of all training texts
  - Say we have 5 topics: Business, Health, Sports, Politics, International
- Find *vocabulary*:
  - Find all words that appear in all training texts.
  - Say we found 2300 words.
- Learning P(Y)
  - P(Y) is a histogram over the 5 topics



Naïve Bayes for Text Classification

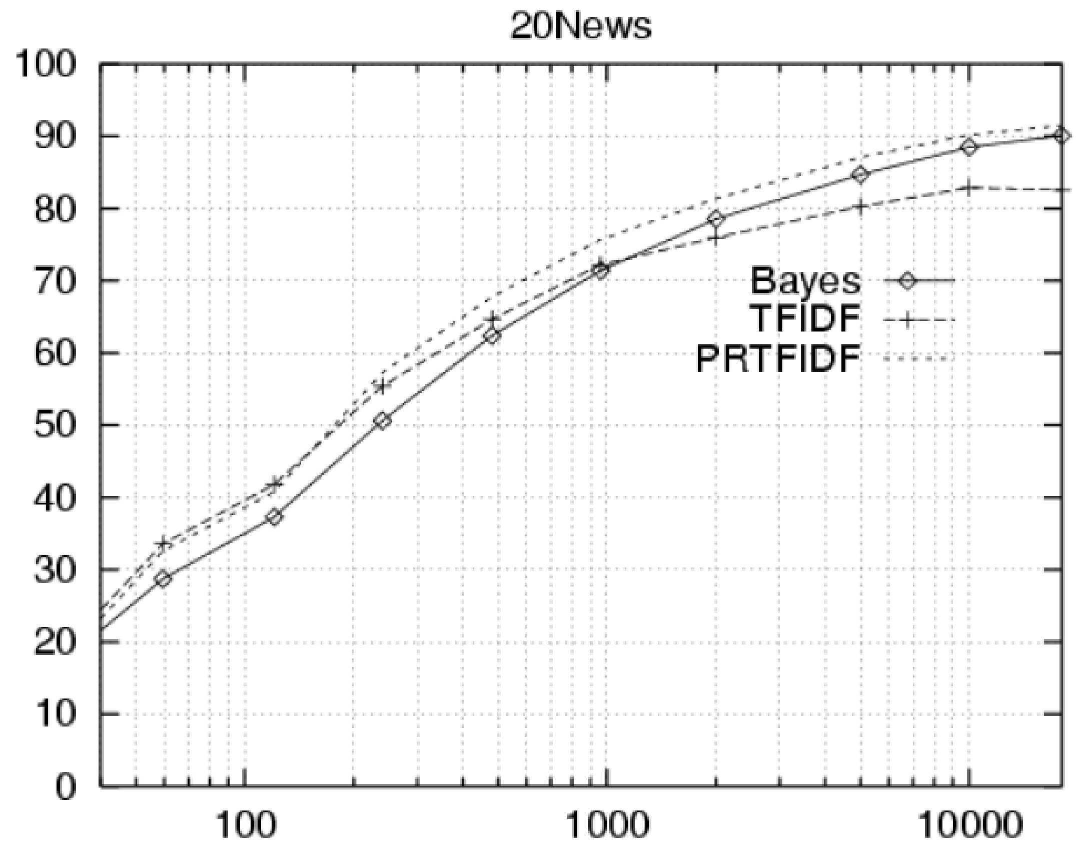
- The observed variables are the words at each position
  - X_1 the first word, X_2 = the second word, …
  - For a given topic y_i, variables X_i are assumed iid from $P(X|Y=y_i)$
- Learning $P(X|Y=y_i)$
  - For each y_i, $P(X|Y=y_i)$ is a histogram over the words
  - 2300 bins in our example
  - Initialize 1 sample in each bin
- Classification:
  - Say the new text has n words
  - The topic is:



$$Y = \arg\max_{y_k} P(y_k) \prod_{i=1}^{n} P(w_i|y_k)$$

Example: 20 Newsgroups Classification

- 20000 documents, 1000 from each newsgroup
- 100 most frequent words removed
- Words occurring at most three times removed
- Vocabulary obtained had 38500 words
- Result: 89% accuracy



Accuracy vs. Training set size (1/3 withheld for test)

Continuous Features

- If X_i is continuous, fit a parametric model for $P(X_i|Y)$
  - E.g. Gaussian

$$P(X_i = x|Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{-\frac{(x-\mu_{ik})^2}{2\sigma_{ik}^2}}$$

  - Sometimes assume σ_{ik} is
    - Same for all features in X: $\sigma_{ik}=\sigma_k$
    - Independent of Y: $\sigma_{ik}=\sigma_i$
    - Constant: $\sigma_{ik}=\sigma$
  - Parameter estimation:
    - Collect all values of X_i^j for which $Y^j=k$
    - Compute mean and variance of these samples

Conclusions

- Naïve Bayes
  - Simple generative model
  - Assumes independence of predictors
  - Effective for text classification

- Learning for Naïve Bayes
  - Non-parametric models:
    - Histograms, Parzen windows
  - Parametric models:
    - Gaussians, mixture of Gaussians