

Lecture 3 – Naïve Bayes, Maximum Entropy and Text Classification

COSI 134



Conditional Parameterization

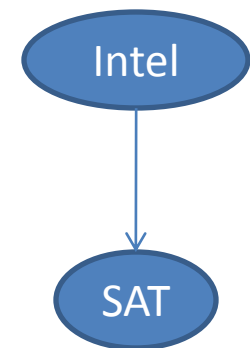
- Two RVs: Intelligence(I) and SAT(S)
- $\text{Val}(I) = \{\text{High}, \text{Low}\}$, $\text{Val}(S) = \{\text{High}, \text{Low}\}$
- A possible joint distribution
- Can describe using chain rule as

$$P(I,S) = P(I)P(S|I)$$

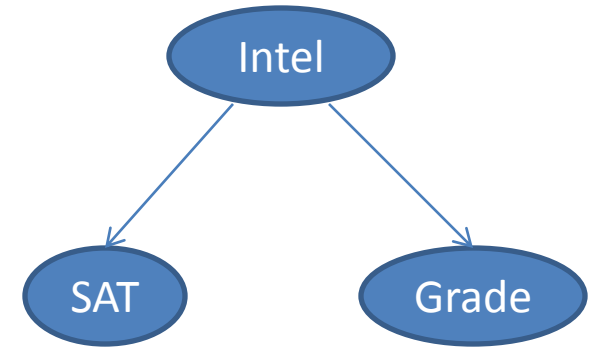
| I | S | P(I,S) |
|------|------|--------|
| Low | Low | 0.665 |
| Low | High | 0.035 |
| High | Low | 0.06 |
| High | High | 0.24 |

| P(I=Low) | P(I=High) |
|----------|-----------|
| 0.7 | 0.3 |

| P(S I) | S=Low | S=High |
|--------|-------|--------|
| I=Low | 0.95 | 0.05 |
| I=High | 0.2 | 0.8 |



Conditional Independence



- Assume another RV, Grade(G)

- Grade in some course

- $\text{Val}(G) = \{\text{High, Medium, Low}\}$

- Might assume that G is conditionally independent of S given I

$$P(G | I, S) = P(G | I)$$

- Then: $P(I, S, G) = P(S, G | I)P(I)$

By cond. indep. $P(S, G | I) = P(S | I)P(G | I)$

So, $P(I, S, G) = P(S | I)P(G | I)P(I)$

- Another CPT for $P(G | I)$

- More compact than full joint

- Possible to update joint with new information

| $P(G I)$ | G=High | G=Med | G=Low |
|------------|--------|-------|-------|
| I=Low | 0.2 | 0.34 | 0.46 |
| I=High | 0.74 | 0.17 | 0.09 |

Statistical Modeling

- Four Questions

- 1) What is the form of the **model**?
 - What random variables? How are probabilities computed? What distributions? What parameters?
- 2) Given a set of data (items from the sample space), how is the **likelihood** of that data computed, for the given model structure and parameter values?
- 3) Given a likelihood function, how are the “optimal” parameters **estimated** given a set of data?
- 4) Given a model form and a set of induced parameter values, how is **inference** performed in the model to make predictions/ask queries

Random Variable Distributions

- Bernoulli Distribution

- Outcome is success (1) or failure (0)
- Success with probability p
- Probability mass function $P(X=1) = 1 - P(X=0) = p$

- Categorical Distribution

- Outcome is one of a finite number of categories
- Probability mass function $P(X = x_i) = p_i$ $\sum_{i=1}^n p_i = 1$

- Binomial Distribution is a series of Bernoulli trials

- Multinomial Distribution is a series of Categorical trials

Naïve Bayes

- Very simple, but effective probabilistic classifier

$$p(y | x_1, \dots, x_n) = \frac{p(y, x_1, \dots, x_n)}{p(x_1, \dots, x_n)} = \frac{p(x_1, \dots, x_n | y)p(y)}{p(x_1, \dots, x_n)}$$

- But – how do we calculate $p(x_1, \dots, x_n | y)$

- Naïve Bayes Assumption: $p(x_1, \dots, x_n | y) = \prod_{i=1}^n p(x_i | y)$

- Each observed variable is assumed to be independent of each other given the class

Naïve Bayes Inference

- First, note that to use the model in most settings, we do not need to explicitly compute

$$\frac{p(x_1, \dots, x_n | y)p(y)}{p(x_1, \dots, x_n)}$$

- Denominator can be ignored since the data are given and the same across all y
- We are interested in

$$\begin{aligned} \arg \max_y (p(y | x_1, \dots, x_n)) &= \arg \max_y \frac{p(x_1, \dots, x_n | y)p(y)}{p(x_1, \dots, x_n)} \\ &= \arg \max_y p(x_1, \dots, x_n | y)p(y) \end{aligned}$$

Example: Document Classification

DOCUMENTS:

To finance extra spending on Labour's policies, such as education, Mr. Brown announced that the Treasury would collect 30 billion pounds by selling national assets like the Tote as well as government shares in British Energy and the

FINANCE

England have won the third Test at Mumbai by 212 runs and secured a share of the series in which few observers, if any, gave them hope of avoiding defeat. Set 313 to win, India folded to 100 all out an hour and a half into the afternoon session, with their ...

SPORTS

Classify documents based on their vocabulary.

$$p(\text{class} = C \mid w_{\text{Brown}} = 1, w_{\text{finance}} = 1, w_{\text{spending}} = 1, w_{\text{Treasury}} = 1, \dots)$$

Observed Variables in NB

- The X variables in $p(x_1, \dots, x_n | y)$
- Bernoulli model introduces a set of Bernoulli RVs, one for each item in our vocabulary, such that $X_w = 1$ iff w appears in the document
- The multinomial model introduces an RV for each position in a document. The RV is multinomial, ranging over the vocabulary
 - E.g. $X_1 = \textit{England}, X_2 = \textit{have}, X_3 = \textit{won}$
 - But, we'd like positional independence

$$p(X_i = \textit{England} | C) = p(X_j = \textit{England} | C)$$

Generative Story

- Bernoulli Case

- 1) Generate a document class from $p(y)$
- 2) Generate an indicator variable X_i for each vocabulary item
- 3) Generate words according to which $X_i = 1$

- Multinomial Case

- 1) Generate a document class from $p(y)$
- 2) For each position k , generate a word from $p(X_k = w | C)$
- 3) Do this for all positions in document
 - Note that true generative model would require modeling document length

Estimation

- Maximum likelihood estimation $p(D | \theta) = \prod_{i=1}^n p(x^{(i)}, y^{(i)})$

$$\log p(D_{1:n} | \theta) = \log \prod_{k=1}^n p(x^{(k)}, y^{(k)}) = \sum_{k=1}^n \log p(x^{(k)}, y^{(k)}) = \sum_{k=1}^n \log p(x^{(k)} | y^{(k)}) + \log p(y^{(k)})$$

- We need to find estimates for $p(y)$
- And for class conditional posteriors $p(x_i | y)$
- That MAXIMIZE the likelihood

Estimation Cont.

$c(x, y)$ = # documents of class y that x occurs in

$c'(x, y)$ = # of times x occurs across documents of class y

- Bernoulli ML estimate $p(x_i | y) = \frac{c(x_i, y)}{c(y)}$
- Multinomial ML estimate $p(x_i | y) = \frac{c'(x_i, y)}{\sum_j c'(x_j, y)}$
- Class prior ML estimate $p(y) = \frac{c(y)}{\sum_{y'} c(y')}$

Smoothing

- Estimates can be problematic with small amounts of data
- Other estimates can be more reliable

- Laplace smoothing

$$p(x_i | y) = \frac{c(x_i, y) + 1}{c(y) + 2}$$

- Generalized Laplace smoothing $p(x_i = v_j | y) = \frac{c(x_i, v_j, y) + 1}{c(y) + s_i}$

- Where $s_i = |\text{Val}(x_i)|$

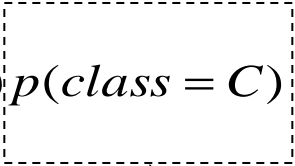
Document Classification with NB

$$p(\text{class} = C \mid w_{\text{Brown}} = 1, w_{\text{finance}} = 1, w_{\text{spending}} = 1, w_{\text{Treasury}} = 1, \dots)$$

Is proportional to:

$$p(w_{\text{Brown}} = 1, w_{\text{finance}} = 1, w_{\text{spending}} = 1, w_{\text{Treasury}} = 1 \mid \text{class} = C) p(\text{class} = C)$$

$$p(w_{\text{Brown}} = 1, w_{\text{finance}} = 1, w_{\text{spending}} = 1, w_{\text{Treasury}} = 1, \dots \mid \text{class} = C) = \\ p(w_{\text{Brown}} = 1 \mid \text{class} = C) p(w_{\text{finance}} = 1 \mid \text{class} = C) p(w_{\text{spending}} = 1 \mid \text{class} = C) \dots$$



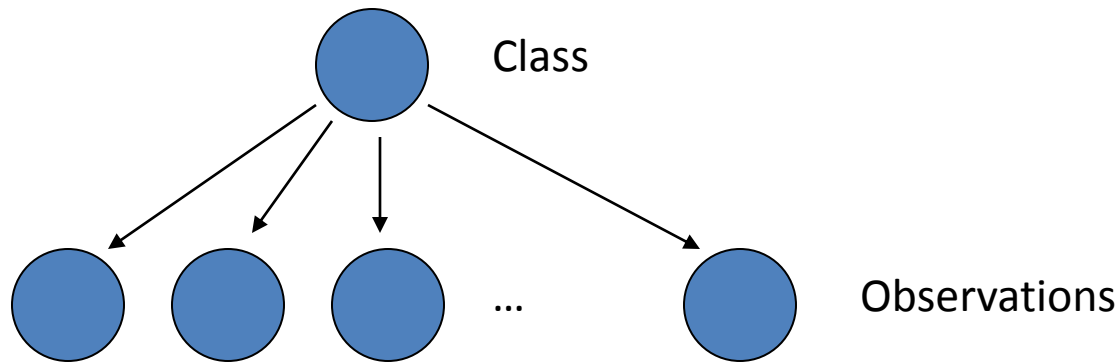
Class prior probability is just the frequency of the class in the training data.

Note that the model assumes each word in a document is independent, *given the class* of the document.

Clearly, this assumption is wrong. However, the classifier still performs well in practice.

Preview of Graphical Models

- Naïve Bayes is a simple model
- Strong conditional independence assumptions



- Graphical models allow us to determine/specify conditional independence assumptions
- Facilitate development of **algorithms** for learning and inference

Motivation for Conditional Model

- Strong independence assumptions in NB
- Results in poorly calibrated posterior probabilities

- Also, NB is **generative**

- It models the joint distribution $p(y | x_1, \dots, x_n) = \frac{p(y, x_1, \dots, x_n)}{p(x_1, \dots, x_n)}$
- It can generate the observed data (e.g. given a class)
- AND make predictions about the class given the data
- We usually only care about making predictions
- Modeling “power” is used to properly generate the data