Bayesian Networks and Naive Bayes
J&M Chapter 4; Collins EM

CSE 597: Natural Language Processing

Outline

- Bayesian Networks and Probability (Some review)
- Naive Bayes
- Smoothing for NB
- Estimating NB from Unlabeled Data: Expectation Maximization

Bayesian Networks and Naive Bayes
J&M Chapter 4; Collins EM

CSE 597: Natural Language Processing

Bayesian Networks and Probability

Probability vs. Joint Probability

- Probability of a proposition is the sum of probabilities in all situations:

$$P(\Omega) = \sum_{\omega \in \Omega} P(\omega)$$

- Joint probability of two propositions A, Ω: the probability of their co-occurrence:

$$P(A, \Omega)$$

- Two variables are independent *iff* . . .

Probability vs. Joint Probability

- Probability of a proposition is the sum of probabilities in all situations:

$$P(\Omega) = \sum_{\omega \in \Omega} P(\omega)$$

- Joint probability of two propositions A, Ω: the probability of their co-occurrence:

$$P(A, \Omega)$$

- Two variables are independent *iff* . . .

$$P(A, \Omega) = P(A) \cdot P(\Omega)$$

Probabilities of Elementary Events

- Assumption 1: The set Ω of all possible complete specifications of states of the world is known (power set, all subsets of Ω)

$$\text{power set of } \{x, y, z\} = \{\{x, y, z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x\}, \{y\}, \{z\}, \emptyset\}$$

- Assumption 2: Every $\omega_i \in \Omega$ is assigned a probability $P(\omega_i)$:

  (What range of values will P(ω_i) have?)

  and assuming Ω is finite, then P is the probability distribution on Ω

  (What can we say about the relation of P(ω_i) to P(Ω)?)

Probabilities of Elementary Events

- Assumption 1: The set Ω of all possible complete specifications of states of the world is known (power set, all subsets of Ω)

$$\text{power set of } \{x, y, z\} = \{\{x, y, z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x\}, \{y\}, \{z\}, \emptyset\}$$

- Assumption 2: Every $\omega_i \in \Omega$ is assigned a probability $P(\omega_i)$:

$$0 \leq P(\omega_i) \leq 1$$

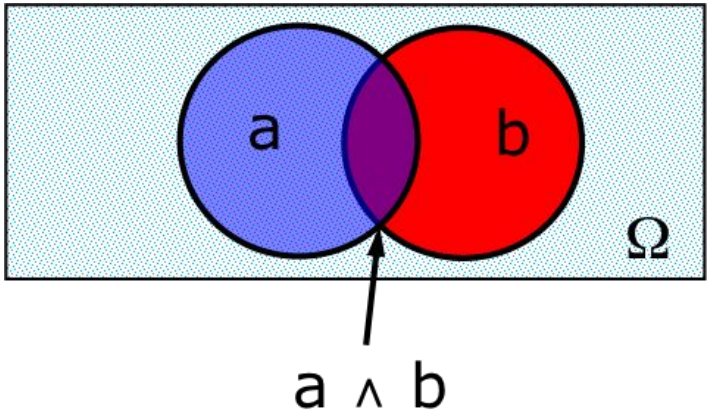
and assuming Ω is finite, then P is the probability distribution on Ω

$$P(\Omega) = \sum_{\omega \in \Omega} P(\omega) = 1$$

Probabilities of Events: Logical truth, AND, OR

For any finite, discrete events a and b:

$$P(\emptyset) = 0 = P(false)$$
$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$
$$\neg P(a) = 1 - P(a)$$



Let the following probabilities hold:

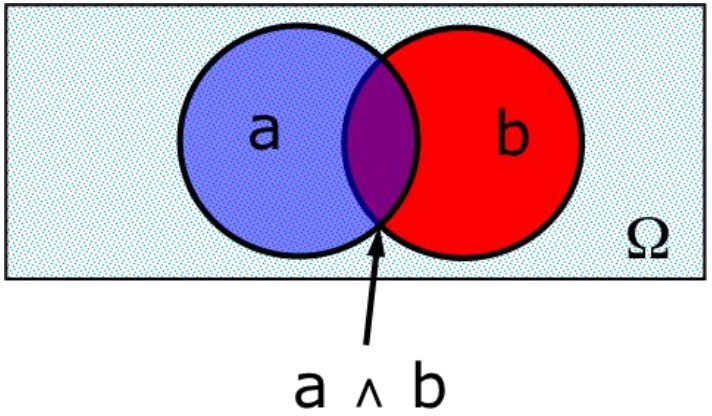
$$P(a) = 0.4; P(b) = 0.3$$

What range of values must $P(a \wedge b)$ have if $a \cap b \neq 0$?

Probabilities of Non-disjoint Events

For any finite, discrete events a and b:

$$P(\emptyset) = 0 = P(false)$$
$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$
$$\neg P(a) = 1 - P(a)$$



Let the following probabilities hold:

$$P(a) = 0.4; P(b) = 0.3$$

What range of values must $P(a \wedge b)$ have if $a \cap b \neq 0$?

$$0 < P(a \wedge b) \leq 0.3$$

Independence Examples: flipping 2 coins

- Two events A and B are independent if $P(A \cap B) = P(A)\,P(B)$
- Two events are dependent if they are not independent
- Examples of coin tosses: independent pairs of events or not?
  - $\Omega = \{HH, HT, TH, TT\}$, and $\forall \omega \in \Omega, P(\omega) = 1/|\Omega| = 0.25$
  - Event A = 2nd flip is H = {HH, HT}, P(A) = 0.5
  - Event B = 1st flip is H ={HH, TH}, P(B) = 0.5
  - $P(A \cap B)$ = **?**
  - Event C = flip contains a T = {HT, TH, TT}
  - $P(A \cap C)$ = **?**

Independence Examples: flipping 2 coins

- Two events A and B are independent if $P(A \cap B) = P(A)\,P(B)$
- Two events are dependent if they are not independent
- Examples of coin tosses: independent pairs of events or not?
  - $\Omega = \{HH, HT, TH, TT\}$, and $\forall \omega \in \Omega, P(\omega) = 1/|\Omega| = 0.25$
  - Event A = 2nd flip is H = {HH, HT}, P(A) = 0.5
  - Event B = 1st flip is H ={HH, TH}, P(B) = 0.5
  - $P(A \cap B) = P(HH) = 0.25 = P(A)P(B)$ *Independent: 0.5 x 0.5 = 0.25*
  - Event C = flip contains a T = {HT, TH, TT}
  - $P(A \cap C) = P(HT) = 0.25 \neq P(A)P(C)$ *Dependent: 0.5 x 0.75 ≠ 0.25*

Conditional Probability

- To marginalize out a subset of probabilities for some situation Ω in P(A, Ω), sum over the cases of A:

$$P(\Omega) = \sum_{\alpha \in A} P(A, \Omega)$$

- A conditional probability conditioned on the set of situations Ω normalizes the joint distribution by P(Ω)

$$P(A|\Omega) = \frac{P(A,\Omega)}{P(\Omega)}$$

$$P(A, \Omega) = P(A|\Omega)P(\Omega)$$

Conditional Probability Example

- A family has two children. What is the probability that both are boys, given that at least one is a boy?

Conditional Probability Example

- A family has two children. What is the probability that both are boys, given that at least one is a boy?
- Hint 1: you do not need to know P(Boy) or P(¬Boy)

Conditional Probability Example

- A family has two children. What is the probability that both are boys, given that at least one is a boy?
- Hint 1: you do not need to know P(B) or P(¬B)
- Hint 2: you need to consider all possibilities

Conditional Probability Example

- A family has two children. What is the probability that both are boys, given that at least one is a boy?
- Hint 1: you do not need to know P(B) or P(¬B)
- Hint 2: you need to consider all possibilities
- P(BB|GB ∨ BG ∨ BB) = 1/3

Derivation of Bayes Theorem

- Given P(A), P(Ω) and P(Ω|A), find P(A|Ω)

- Derivation
  - $P(A \cap \Omega) = P(\Omega \cap A)$ *Commutativity*
  - $P(A|\Omega)P(\Omega) = P(\Omega|A)P(A)$ *Definition of conditional probability*
  - $P(A|\Omega) = P(\Omega|A)P(A)/P(\Omega)$ *Simplification*

Chain Rule

- Rewrite conditional probability $P(A|\Omega) = P(A,\Omega)/P(\Omega)$ as $P(A,\Omega) = P(A|\Omega)P(\Omega)$
- More variables

  $P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)$

- General form

  $P(X_1,X_2,\ldots,X_n)=P(X_1)P(X_2|X_1)\ldots P(X_n|X_1,X_2,\ldots X_{n-1})$

- Recall the Markov Assumption: A type of conditional independence pertaining to time or sequence

Full Joint Probability Distributions

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

- Consider three random boolean variables: Toothache, Cavity, Catch
  - Gives a 2 x 2 x 2 = 8 sample space of distinct events
    - Four cavity and four ¬cavity
    - Four toothache and four ¬toothache
    - Four catch and four ¬catch
  - Sum of probabilities in the full joint distribution = 1

Full Joint Probability Distributions

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

- Full joint probability distribution can be used to answer questions about probabilities of all possible events
  - Six ways to have (toothache ∨ cavity)

$$P(cavity \vee toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

  - Marginal probability of cavity

$$P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.20$$

Marginalization and Conditioning

- Marginalizing (summing out) for a variable sums over all the values of another variable in the joint distribution

$$P(Y) = \sum_z P(Y, Z = z)$$

- Conditioning, derived from applying the product rule to the rule for marginalizing

$$P(Y) = \sum_z P(Y|Z)P(Z)$$

Illustration of Conditional Probabilities

- Probability of a cavity conditioned on toothache

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)}$$
$$= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6.$$

- Probability of not having a cavity conditioned on toothache

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$
$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4.$$

Joint Probability Distribution with Independence

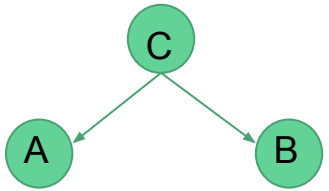
Given three binary random variables A, B, C, full joint distribution P(A,B,C) has 2 x 2 x 2-1 = 7 independent values

	+A		-A	
	+B	-B	+B	-B
+C	+A+B+C	+A-B+C	-A+B+C	-A-B+C
-C	+A+B-C	+A-B-C	-A+B-C	-A-B-C

Conditional Independence Simplifies Things

A and B are conditionally independent given C iff

- P(A|B,C) = P(A|C)
- P(B|A,C) = P(B|C)
- P(A ⋀ B|C) = P(A|C) P(B|C)



With conditional independence of A, B given C:

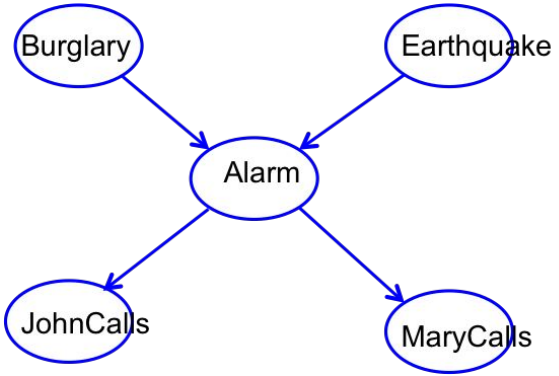
- P(A,B,C) = P(A|B,C)P(B|C)P(C) (chain rule)
  = P(A|C)P(B|C)P(C) (conditional independence)
  =(2 x 2 + 2) - 1 = 5 independent values:

Bayesian Networks

- A simple graphical notation for joint probability distributions
  - Represents conditional independence relations
  - Leads to **compact** representations of joint probability distributions
- Syntax of BNs: Directed Acyclic Graphs (DAGs)
  - A set of nodes, one per random variable
  - Edges from each conditioning (evidence) node to its children
  - A conditional probability distribution for each node X_i given its parents $P(X_i|Parents(X_i))$
- In simple cases (e.g., binary variables), conditional distributions for each node are represented as conditional probability tables (CPTs)

Bayesian Networks for Causal Relations

- Problem: I'm at work, my neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. They usually call if my burglar alarm goes off, and occasionally otherwise. Sometimes it's set off by minor earthquakes. Is there a burglar?
- Variables: Burglary, Earthquake, Alarm, JohnCalls, Mary Calls
- Network topology as "causal" knowledge:
  - P(Alarm|Burglary, Earthquake)
  - P(JohnCalls, MaryCalls|Alarm)

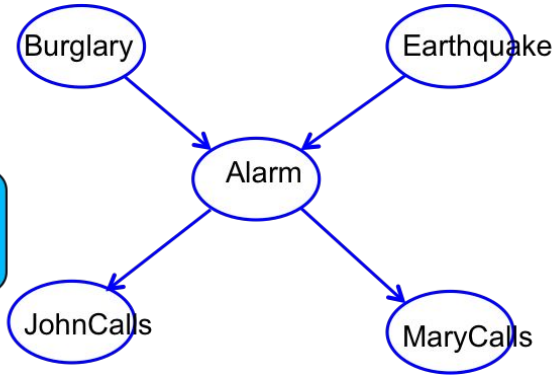


Semantics

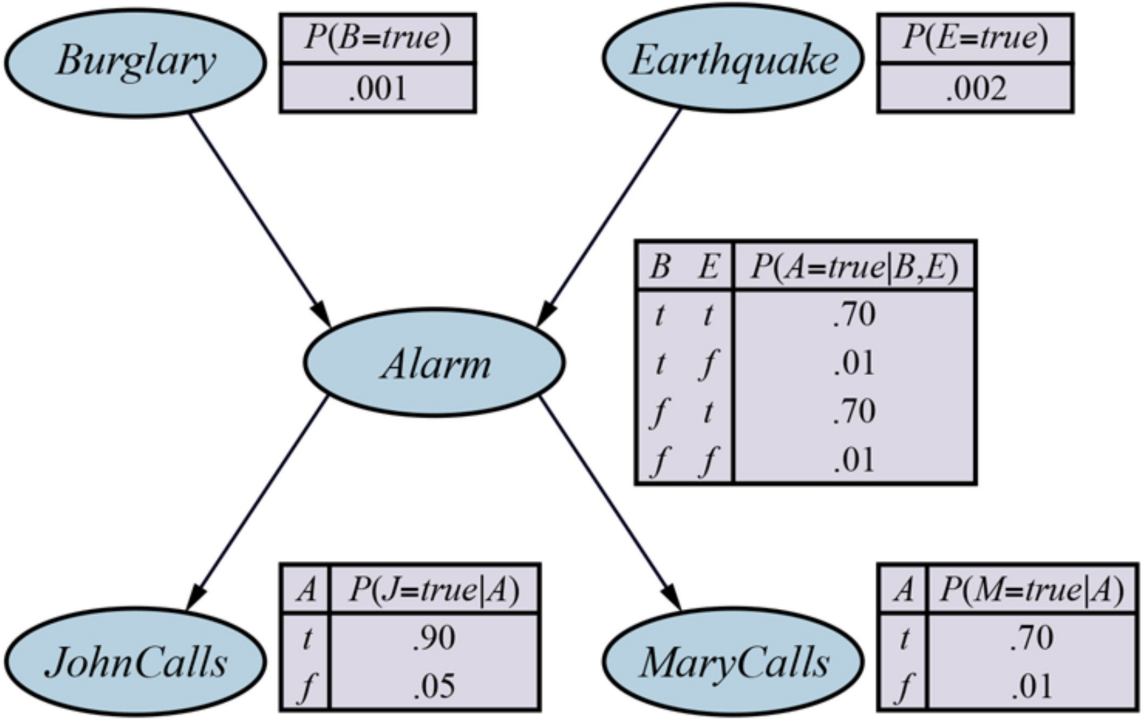
- The full joint probability definition is defined as the product of the local conditional distributions
- The NB model is simpler, as we will see

$$P(x_1,\ldots,x_n)=\prod_{i=1}^{n}P(x_i \mid parents(X_i))$$

Generalizes Naïve Bayes



Belief Network with Conditional Probability Tables



Construction Algorithm for Bayesian Network

1. Choose an ordering of variables X_1, ..., X_n
2. For i = 1 to n
   a. Add X_i to the network
   b. Select parents from X_1, ..., X_{i-1} such that $P(X_i|Parents(X_i)) = P(X_i|X_1,...,X_{i-1})$
3. Alternative networks can be constructed for the same set of random variables
   a. Causal networks, parents are causes
   b. Diagnostic networks, parents are symptoms
4. The networks are smallest when causes are parents of effects

Bayesian Networks and Naive Bayes
J&M Chapter 4; Collins EM

CSE 597: Natural Language Processing

Naive Bayes

Full Joint Probability Distribution

	Topic	
Words	Archery	¬ Archery
quiver	0.01	0.01
¬ *quiver*	0.09	0.89

- Probabilities of all events in Ω must sum to 1
- Consider two types of random variables:
  - Variables X_i for each **word** in the vocabulary: *arrow, …, quiver, …, wicket*
  - Variables Y_j for each **topic** in a set of documents about sports played in Britain: Archery, Fencing, Croquet, Lawn Bowls
- Naive Bayes classifiers: What are the parameters of the X_i for a given Y_j

Conditional Independence Assumption for **NB**

- Given a **single** parent variable Y and d children X_d:

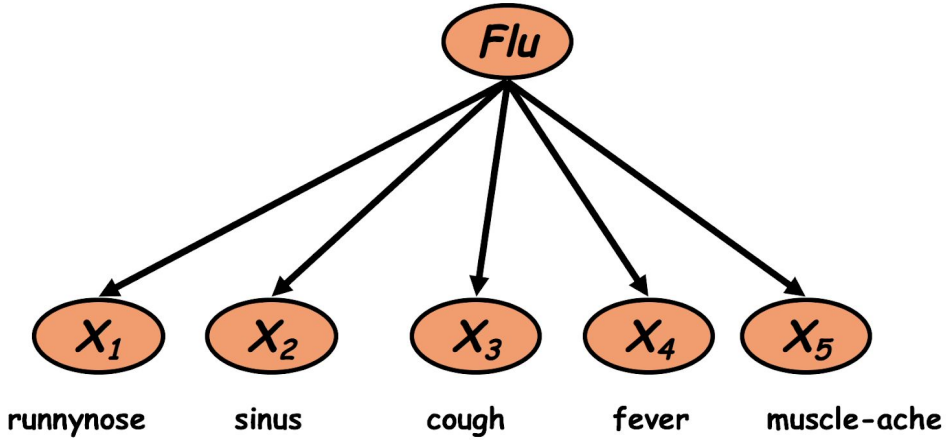
$$P(Y = y, X_1 = x_1, X_2 = x_2, \ldots X_d = x_d)$$

$$= \quad P(Y = y) \prod_{j=1}^{d} P(X_j = x_j | Y = y)$$

- **Multiple** parent variables Y_j corresponds to ***multinomial NB***

Naive Bayes Graph (Single Parent)

- All effects assumed to be conditionally independent, given the cause
- Total number of parameters is linear in number of effects
- ***Naive*** because can work well when conditional independence does not hold



Maximum Likelihood Estimation

- The statistical inference problem is to estimate the parameters θ given observations x_i for the classes y_j
- Maximum likelihood estimation (MLE) chooses the estimate θ^* that maximizes the likelihood function, i.e.:

$$\theta^* = \operatorname{argmax} L(\theta \mid x, y) = \operatorname{argmax} p_\theta(y \mid \theta, x)$$

- That is, the MLE parameters are those that maximize the probability of the observed data distribution in the classes

Example: Will the tennis match be cancelled?

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example: Will the tennis match be cancelled?

$$x_{outlook} \equiv \{Sunny, Overcast, Rain\}$$

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
				Weak	No
				Strong	No
				Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$y \in \{0, 1\}$ (No, Yes)

Learning phase: MLE Estimates

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temp.	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=*No*
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

- P(Play = yes) = 9/14
- P(Play = no) = 5/14

Test phase: Maximum A Posteriori Decision Rule

- A new instance, for class C and data D:

  x'=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)

$$P(c|d) = \alpha P(d|c)P(c)$$

$$c_{MAP} \equiv \arg\max_{c \in C} P(c|d)$$

P(*Yes* | **x'**): [P(*Sunny* | Yes)P(*Cool* | *Yes*)P(*High* | Yes)P(*Strong* | *Yes*)]P(Play=*Yes*) = 0.0053

(2/9 x 3/9 x 3/9 x 3/9) x 9/14

P(*No* | **x'**): [P(*Sunny* | No) P(*Cool* | No)P(*High* | *No*)P(*Strong* | *No*)]P(Play=*No*) = 0.0206

(3/5 x 1/5 x 4/5 x 3/5) x 5/14

Given the fact P(*Yes* | **x'**) < P(*No* | **x'**), we label **x'** to be "*No*"

Naive Bayes, Formalized

The setting is as follows. Assume we have some training set $(\underline{x}^{(i)}, y^{(i)})$ for $i = 1 \ldots n$, where each $\underline{x}^{(i)}$ is a vector, and each $y^{(i)}$ is in $\{1, 2, \ldots, k\}$.

We will assume throughout that each vector $\underline{x}$ is in the set $\{-1, +1\}^d$ for some integer d specifying the number of "features" in the model. In other words, each component x_j for $j = 1 \ldots d$ can take one of two possible values.

From Collins reading on NB, MLE, EM

Definition 1 (Naive Bayes (NB) Model) *A NB model consists of an integer k specifying the number of possible labels, an integer d specifying the number of attributes, and in addition the following parameters:*

- *A parameter*

$$q(y)$$

for any $y \in \{1 \dots k\}$. The parameter $q(y)$ can be interpreted as the probability of seeing the label y. We have the constraints $q(y) \geq 0$ and $\sum_{y=1}^{k} q(y) = 1$.

- *A parameter*

$$q_j(x|y)$$

for any $j \in \{1 \dots d\}$, $x \in \{-1, +1\}$, $y \in \{1 \dots k\}$. The value for $q_j(x|y)$ can be interpreted as the probability of attribute j taking value x, conditioned on the underlying label being y. We have the constraints that $q_j(x|y) \geq 0$, and for all y, j, $\sum_{x \in \{-1,+1\}} q_j(x|y) = 1$.

From Collins reading on NB, MLE, EM

We then define the probability for any $y, x_1 \ldots x_d$ as

$$p(y, x_1 \ldots x_d) = q(y) \prod_{j=1}^{d} q_j(x_j|y)$$

The next section describes how the parameters can be estimated from training examples. Once the parameters have been estimated, given a new test example $\underline{x} = \langle x_1, x_2, \ldots, x_d \rangle$, the output of the NB classifier is

$$\arg \max_{y \in \{1 \ldots k\}} p(y, x_1 \ldots x_d) = \arg \max_{y \in \{1 \ldots k\}} \left(q(y) \prod_{j=1}^{d} q_j(x_j|y) \right)$$

From Collins reading on NB, MLE, EM

Maximum Likelihood Estimates (MLE) for NB

- MLE of q(y) is frequency of $y^{(i)} = y$ for all training n examples $(x^{(i)}, y^{(i)})$

$$q(y) = \frac{\sum_{i=1}^{n} [[y^{(i)} = y]]}{n} = \frac{\text{count}(y)}{n}$$

- MLE of $q_j(x|y)$ is the proportion of times the j'th component of the vector $x^{(i)} = x$ given the label y

$$q_j(x|y) = \frac{\sum_{i=1}^{n} [[y^{(i)} = y \text{ and } x_j^{(i)} = x]]}{\sum_{i=1}^{n} [[y^{(i)} = y]]} = \frac{\text{count}_j(x|y)}{\text{count}(y)}$$

From Collins reading on NB, MLE, EM

Log-likelihood Function for MLE Estimation

Given the training set $(x^{(i)}, y^{(i)})$ for $i = 1 \ldots n$:

$$
\begin{aligned}
L(\underline{\theta}) &= \sum_{i=1}^{n} \log p(x^{(i)}, y^{(i)}) \\
&= \sum_{i=1}^{n} \log \left(q(y^{(i)}) \prod_{j=1}^{d} q_j(x_j^{(i)} | y^{(i)}) \right) \\
&= \sum_{i=1}^{n} \log q(y^{(i)}) + \sum_{i=1}^{n} \log \left(\prod_{j=1}^{d} q_j(x_j^{(i)} | y^{(i)}) \right) \\
&= \sum_{i=1}^{n} \log q(y^{(i)}) + \sum_{i=1}^{n} \sum_{j=1}^{d} \log q_j(x_j^{(i)} | y^{(i)})
\end{aligned}
$$

From Collins reading on NB, MLE, EM

Definition 2 (ML Estimates for Naive Bayes Models) *Assume a training set* $(x^{(i)}, y^{(i)})$ *for* $i \in \{1 \ldots n\}$. *The maximum-likelihood estimates are then the parameter values* $q(y)$ *for* $y \in \{1 \ldots k\}$, $q_j(x|y)$ *for* $j \in \{1 \ldots d\}$, $y \in \{1 \ldots k\}$, $x \in \{-1, +1\}$ *that maximize*

$$L(\underline{\theta}) = \sum_{i=1}^{n} \log q(y^{(i)}) + \sum_{i=1}^{n} \sum_{j=1}^{d} \log q_j(x_j^{(i)}|y^{(i)})$$

subject to the following constraints:

1. $q(y) \geq 0$ *for all* $y \in \{1 \ldots k\}$. $\sum_{y=1}^{k} q(y) = 1$.
2. *For all* y, j, x, $q_j(x|y) \geq 0$. *For all* $y \in \{1 \ldots k\}$, *for all* $j \in \{1 \ldots d\}$,

$$\sum_{x \in \{-1,+1\}} q_j(x|y) = 1$$

From Collins reading on NB, MLE, EM

MLE for Multinomial Distributions

Consider the following setting. We have some finite set $\mathcal{Y}$. A *distribution* over the set $\mathcal{Y}$ is a vector q with components q_y for each $y \in \mathcal{Y}$, corresponding to the probability of seeing element y. We define $\mathcal{P}_\mathcal{Y}$ to be the set of all distributions over the set $\mathcal{Y}$: that is,

$$\mathcal{P}_\mathcal{Y} = \{q \in \mathbb{R}^{|\mathcal{Y}|} : \forall y \in \mathcal{Y}, q_y \geq 0; \quad \sum_{y \in \mathcal{Y}} q_y = 1\}$$

In addition, assume that we have some vector c with components c_y for each $y \in \mathcal{Y}$. We will assume that each $c_y \geq 0$. In many cases c_y will correspond to some "count" taken from data: specifically the number of times that we see element y. We also assume that there is at least one $y \in \mathcal{P}_\mathcal{Y}$ such that $c_y > 0$ (i.e., such that c_y is strictly positive).

From Collins reading on NB, MLE, EM

MLE for Multinomial Distributions, *continued*

Definition 3 (ML estimation problem for multinomials) *The input to the problem is a finite set $\mathcal{Y}$, and a weight $c_y \geq 0$ for each $y \in \mathcal{Y}$. The output from the problem is the distribution q^* that solves the following maximization problem:*

$$q^* = \arg\max_{q \in \mathcal{P}_\mathcal{Y}} \sum_{y \in \mathcal{Y}} c_y \log q_y$$

From Collins reading on NB, MLE, EM

MLE for Multinomial Distributions, *continued*

Thus the optimal vector q^* is a distribution (it is a member of the set $\mathcal{P}_{\mathcal{Y}}$), and in addition it maximizes the function $\sum_{y \in \mathcal{Y}} c_y \log q_y$.

We give a theorem that gives a very simple (and intuitive) form for q^*:

Theorem 2 *Consider the problem in definition 3. The vector q^* has components*

$$q_y^* = \frac{c_y}{N}$$

for all $y \in \mathcal{Y}$, where $N = \sum_{y \in \mathcal{Y}} c_y$.

From Collins reading on NB, MLE, EM

Bayesian Networks and Naive Bayes
J&M Chapter 4; Collins EM

CSE 597: Natural Language Processing

Smoothing

MLE Estimates for NB Text Classification: Step 1

- Given a binary text classification C = {pos, neg}
- MLE for the prior probability of pos and neg

$$\hat{P}(Doc_{pos}) = \frac{Count(Doc_{pos})}{Count(Doc_{pos} + Doc_{neg})}$$

$$\hat{P}(Doc_{neg}) = \frac{Count(Doc_{neg})}{Count(Doc_{pos} + Doc_{neg})}$$

MLE Estimates for NB Text Classification: Step 2

- Identify the set of unique words w_i in the documents, V
- Calculate MLEs for the conditional probability of each word $w_i \in V$, given the Class

$$\hat{P}(w_i|pos) = \frac{Count(w_i, Doc_{pos})}{\sum_{w_j \in V} Count(w_j \in Doc_{pos})}$$

$$\hat{P}(w_i|neg) = \frac{Count(w_i, Doc_{neg})}{\sum_{w_j \in V} Count(w_j \in Doc_{neg})}$$

MLE Estimates for ***Missing*** Examples

- What if there were no cases of word w_i in Doc_{neg}? What is the estimate of p(w_i|Doc_{neg})?
- Then log p(x|y) is **undefined**
- And sum of logs is undefined (product of probabilities is zero)
- Solution: **smoothing**

Cromwell's Rule

In **Bayesian** approaches

- Use of **prior probabilities of 0 or 1** should be avoided
- Because: if the prior probability is 0 or 1, then so is the posterior, and no evidence can be considered

Named by statistician Dennis Lindley in reference to Oliver Cromwell, based on well-known quote before the General Assembly of the Church of Scotland in 1650:

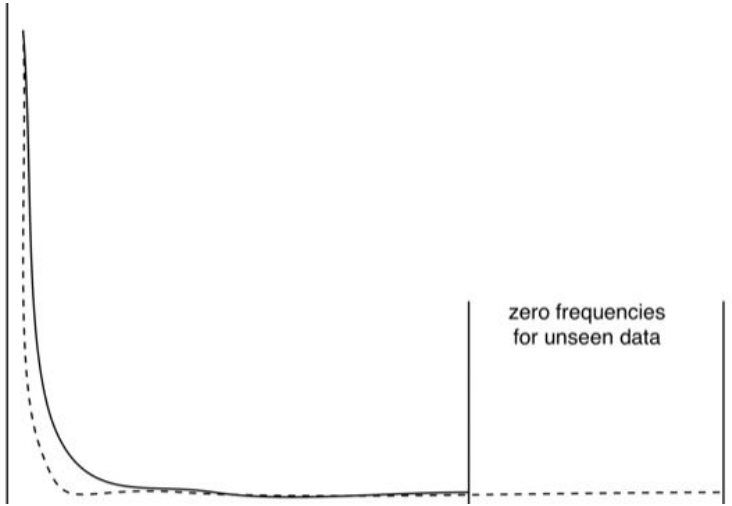
> *I beseech you, in the bowels of Christ,* ***think it possible that you may be mistaken****.*

How to Apply Cromwell's Rule

- Assume we know how many types *never* occur in the data
- Steal probability mass from types that occur at least once
- Distribute this probability mass over the types that never occur

Or, emulate the fictional Robin Hood

- Redistribute the wealth
- Steal from the rich to give to the poor



Smoothing

Add-One La Place Smoothing

- Add **pseudo counts** for **unobserved class members** where the missing word could have occurred (e.g., if only you had more data):
  - Add 1 to the numerator in the conditional probabilities for every word
  - Increment the denominator total count of words in the class by the size of the vocabulary (the class size is expanded by all the pseudo counts)

$$\hat{P}(x_i|c_j) = \frac{Count(x_i, c_j) + 1}{\sum_{x_j \in V}(Count(x_j, c_j) + 1)}$$

$$= \frac{Count(x_i, c_j) + 1}{(\sum_{x_j \in V} Count(x_j, c_j)) + |V|}$$

More General Form of La Place

- Add α, a **much smaller number than 1**

$$\hat{P}(X_i = x_i | C = c_j) = \frac{Count(X_i = x_i, C = C_j) + \alpha}{Count(C = c_j) + \alpha |X_i|}$$

Pros and Cons of La Place Smoothing

- Pro
  - Very simple technique
  - Addresses the key idea that smoothing compensates for not having enough data: Cromwell's rule
- Cons:
  - Probability of frequent words is underestimated
  - Probability of rare (or unseen) words is overestimated
  - Therefore, too much probability mass is shifted towards unseen words
  - All unseen words are smoothed in the same way
- Many more sophisticated methods for smoothing exist; for this class use La Place

Motivation for Calculation in Log Space (Again)

- Smoothing, especially with low α, leads to values close to 0
- Multiplying lots of probabilities, which are between 0 and 1 by definition, can result in floating-point **underflow**
- Mitigation: calculate in log space
  - Given that $\log(xy) = \log(x) + \log(y)$: perform all computations by **summing logs of probabilities** rather than multiplying probabilities

Example: Sentiment Classification, with Smoothing

- Application: reviews of CMPSC 442 at PSU
  - Should the student take the class?
  - How should the student interpret a single review?
- Classes: {Positive, Negative}
- Data: $|V| = 25; |V_{neg}| = 17; |V_{pos}| = 12$

Neg	Just plain boring
Neg	Entirely predictable and lacking in content
Neg	No interesting information and very little relevance to my life
Pos	Amazing relevance to my life
Pos	The best class I've ever taken

Example adapted from Jurafsky & Martin, Speech & Language Processing, Website for 3rd draft, Chapter 4, Naive Bayes and Sentiment Classification

Predicting the Sentiment from New Evidence

- New review: "*a somewhat predictable class*"
- Priors: $P(neg) = \frac{3}{5} \quad P(pos) = \frac{2}{5}$
- Conditional probabilities for previously observed words, with smoothing

$$P(\text{predictable}|-) = \frac{1+1}{17+25} \qquad P(\text{predictable}|+) = \frac{0+1}{12+25}$$

$$P(\text{class}|-) = \frac{0+1}{17+25} \qquad P(\text{class}|+) = \frac{1+1}{12+25}$$

- Maximum probability hypothesis: negative

$$P(-)P(predictable, class|-) = \frac{3}{5} \times \frac{2 \times 1}{42^2} = 6.8 \times 10^{-4}$$

$$P(+)P(predictable, class|+) = \frac{2}{5} \times \frac{2 \times 1}{37^2} = 5.8 \times 10^{-4}$$

Bayesian Networks and Naive Bayes
J&M Chapter 4; Collins EM

CSE 597: Natural Language Processing

Expectation Maximization

Naive Bayes for Unlabeled Data

Now consider the following setting. We have a training set consisting of vectors $x^{(i)}$ for $i = 1 \ldots n$. As before, each $x^{(i)}$ is a vector with components $x_j^{(i)}$ for $j \in \{1 \ldots d\}$, where each component can take either the value -1 or $+1$. In other words, our training set *does not have any labels*.

$$
\begin{aligned}
x_1 &= +1 \text{ if the document contains the word } \textit{Obama}, -1 \text{ otherwise} \\
x_2 &= +1 \text{ if the document contains the word } \textit{McCain}, -1 \text{ otherwise} \\
x_3 &= +1 \text{ if the document contains the word } \textit{Giants}, -1 \text{ otherwise} \\
x_4 &= +1 \text{ if the document contains the word } \textit{Patriots}, -1 \text{ otherwise}
\end{aligned}
$$

$$
\begin{aligned}
\underline{x}^{(1)} &= \langle +1, +1, -1, -1 \rangle \\
\underline{x}^{(2)} &= \langle -1, -1, +1, +1 \rangle \\
\underline{x}^{(3)} &= \langle +1, +1, -1, -1 \rangle \\
\underline{x}^{(4)} &= \langle -1, -1, +1, +1 \rangle \\
\underline{x}^{(5)} &= \langle -1, -1, +1, +1 \rangle
\end{aligned}
$$

From Collins reading on NB, MLE, EM

Naive Bayes for Unlabeled Data, *continued*

We now describe the parameter estimation method for Naive Bayes when the labels $y^{(i)}$ for $i \in \{1 \ldots n\}$ are missing. The first key insight is that for any example $\underline{x}$, the probability of that example under a NB model can be calculated by marginalizing out the labels:

$$p(\underline{x}) = \sum_{y=1}^{k} p(\underline{x}, y) = \sum_{y=1}^{k} \left(q(y) \prod_{j=1}^{d} q_j(x_j|y) \right)$$

From Collins reading on NB, MLE, EM

Definition 4 (ML Estimates for Naive Bayes Models with Missing Labels) *Assume a training set* $(x^{(i)})$ *for* $i \in \{1 \dots n\}$. *The maximum-likelihood estimates are then the parameter values* $q(y)$ *for* $y \in \{1 \dots k\}$, $q_j(x|y)$ *for* $j\{1 \dots d\}$, $y \in \{1 \dots k\}$, $x \in \{-1, +1\}$ *that maximize*

$$L(\underline{\theta}) = \sum_{i=1}^{n} \log \sum_{y=1}^{k} \left(q(y) \prod_{j=1}^{d} q_j(x_j^{(i)}|y) \right) \tag{10}$$

subject to the following constraints:

1. $q(y) \geq 0$ *for all* $y \in \{1 \dots k\}$. $\sum_{y=1}^{k} q(y) = 1$.

Same as for MLE

2. *For all* y, j, x, $q_j(x|y) \geq 0$. *For all* $y \in \{1 \dots k\}$, *for all* $j \in \{1 \dots d\}$,

$$\sum_{x \in \{-1,+1\}} q_j(x|y) = 1$$

From Collins reading on NB, MLE, EM

Goal: Find The Parameters θ That Maximize The Likelihood Of The Data

- The log-likelihood function is then

$$L(\underline{\theta}) = \sum_{i=1}^{n} \log p(x^{(i)}; \underline{\theta}) = \sum_{i=1}^{n} \log \sum_{y \in \mathcal{Y}} p(x^{(i)}, y; \underline{\theta})$$

- The maximum likelihood estimates are

$$\underline{\theta}^* = \arg\max_{\underline{\theta} \in \Omega} L(\underline{\theta})$$

From Collins reading on NB, MLE, EM

The EM algorithm is an iterative algorithm that defines parameter settings $\underline{\theta}^0, \underline{\theta}^1, \ldots, \underline{\theta}^T$ (again, see figure 2). The algorithm is driven by the updates

$$\underline{\theta}^t = \arg\max_{\underline{\theta} \in \Omega} Q(\underline{\theta}, \underline{\theta}^{t-1})$$

for $t = 1 \ldots T$. The function $Q(\underline{\theta}, \underline{\theta}^{t-1})$ is defined as

$$Q(\underline{\theta}, \underline{\theta}^{t-1}) = \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} \delta(y|i) \log p(x^{(i)}, y; \underline{\theta}) \tag{11}$$

where

$$\delta(y|i) = p(y|x^{(i)}; \underline{\theta}^{t-1}) = \frac{p(x^{(i)}, y; \underline{\theta}^{t-1})}{\sum_{y \in \mathcal{Y}} p(x^{(i)}, y; \underline{\theta}^{t-1})}$$

Thus as described before in the EM algorithm for Naive Bayes, the basic idea is to fill in the $\delta(y|i)$ values using the conditional distribution under the previous parameter values (i.e., $\delta(y|i) = p(y|x^{(i)}; \underline{\theta}^{t-1})$).

From Collins reading on NB, MLE, EM

EM Input and Initialization for Naive Bayes (Fig. 2)

Inputs: An integer k specifying the number of classes. Training examples $(x^{(i)})$ for $i = 1 \ldots n$ where each $x^{(i)} \in \{-1, +1\}^d$. A parameter T specifying the number of iterations of the algorithm.

Initialization: Set $q^0(y)$ and $q_j^0(x|y)$ to some initial values (e.g., random values) satisfying the constraints

- $q^0(y) \geq 0$ for all $y \in \{1 \ldots k\}$. $\sum_{y=1}^{k} q^0(y) = 1$.
- For all y, j, x, $q_j^0(x|y) \geq 0$. For all $y \in \{1 \ldots k\}$, for all $j \in \{1 \ldots d\}$,

$$\sum_{x \in \{-1,+1\}} q_j^0(x|y) = 1$$

From Collins reading on NB, MLE, EM

EM Algorithm and Output for Naive Bayes

Algorithm:

For $t = 1 \ldots T$

1. For $i = 1 \ldots n$, for $y = 1 \ldots k$, calculate

$$\delta(y|i) = p(y|\underline{x}^{(i)}; \underline{\theta}^{t-1}) = \frac{q^{t-1}(y) \prod_{j=1}^{d} q_j^{t-1}(x_j^{(i)}|y)}{\sum_{y=1}^{k} q^{t-1}(y) \prod_{j=1}^{d} q_j^{t-1}(x_j^{(i)}|y)}$$

2. Calculate the new parameter values:

$$q^t(y) = \frac{1}{n} \sum_{i=1}^{n} \delta(y|i) \qquad q_j^t(x|y) = \frac{\sum_{i:x_j^{(i)}=x} \delta(y|i)}{\sum_i \delta(y|i)}$$

Output: Parameter values $q^T(y)$ and $q^T(x|y)$.

From Collins reading on NB, MLE, EM

EM Algorithm and Output for Naive Bayes

Algorithm:

For $t = 1 \ldots T$

1. For $i = 1 \ldots n$, for $y = 1 \ldots k$, calculate

Labels

Data

Parameter Priors

Posterior hypothesis

$$\delta(y|i) = p(y|\underline{x}^{(i)}; \underline{\theta}^{t-1}) = \frac{q^{t-1}(y) \prod_{j=1}^{d} q_j^{t-1}(x_j^{(i)}|y)}{\sum_{y=1}^{k} q^{t-1}(y) \prod_{j=1}^{d} q_j^{t-1}(x_j^{(i)}|y)}$$

2. Calculate the new parameter values:

$$q^t(y) = \frac{1}{n} \sum_{i=1}^{n} \delta(y|i) \qquad q_j^t(x|y) = \frac{\sum_{i:x_j^{(i)}=x} \delta(y|i)}{\sum_i \delta(y|i)}$$

Output: Parameter values $q^T(y)$ and $q^T(x|y)$.

From Collins reading on NB, MLE, EM

Summary of topics

- Bayesian networks capture conditional independence relations in joint probability distributions for (somewhat more) efficient inference
- Naive Bayes is a simple type of BN for binary or multinomial classification problems
- Smoothing must be used in NB or other probabilistic models to avoid zero probability events in the training data
- Expectation Maximization (EM) can be used to estimating NB from unlabeled data