EE4211: Data Science for the Internet of Things

Bayesian Data Analysis

Biplab Sikdar

Agenda

- Examples
- Bayesian classification
- Bayesian parameter estimation

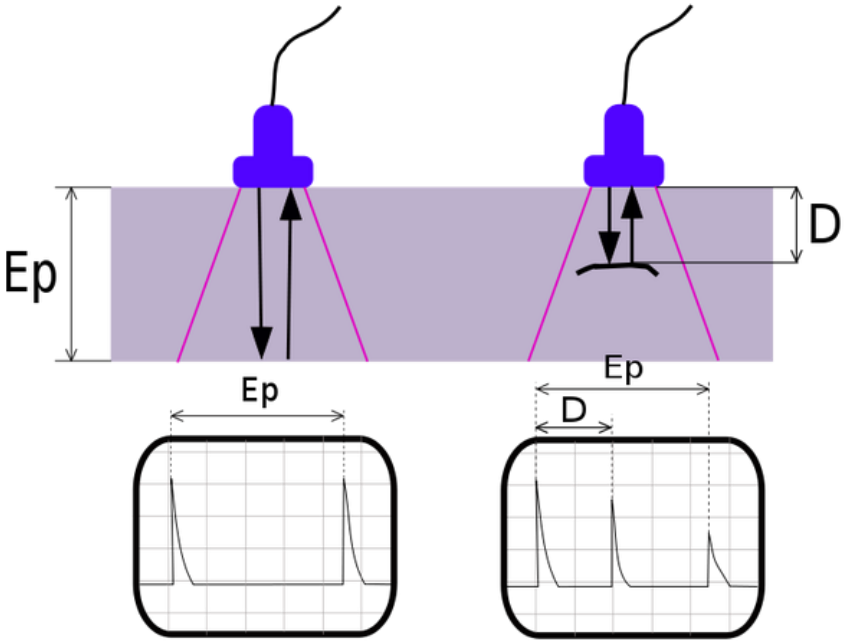
Example 1

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

- <Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=strong>
- Will Roger play tennis?

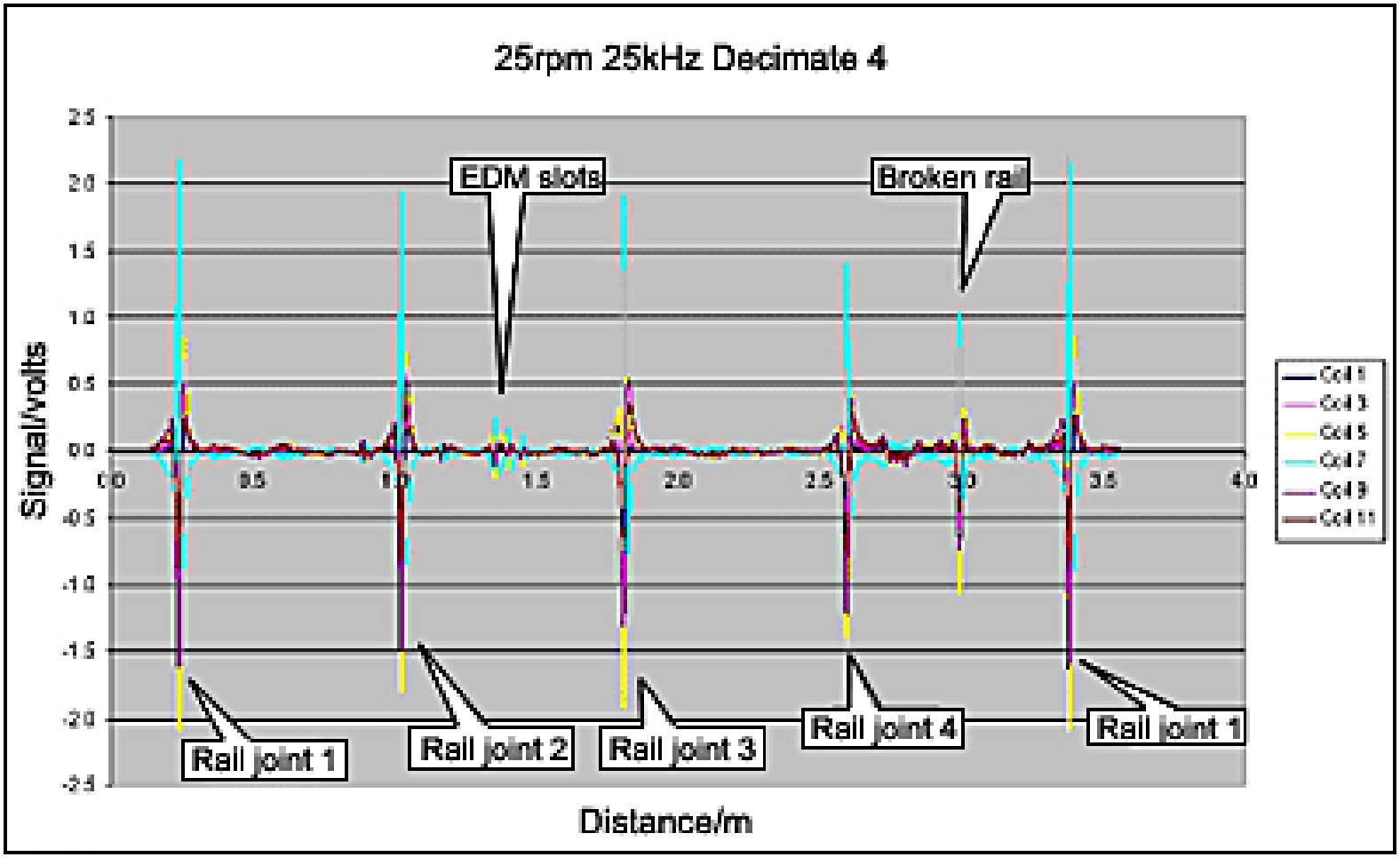
Example: Track Monitoring

Example: Track Monitoring



Example: Track Monitoring

Example: Track Monitoring



Example 3: Document Classification

From: NUS Webmail Alert Notice <mail@nus.edu.sg> Sent: Mon 5/13/2013 1:08 PM
To:
Cc:
Subject: Important Notification

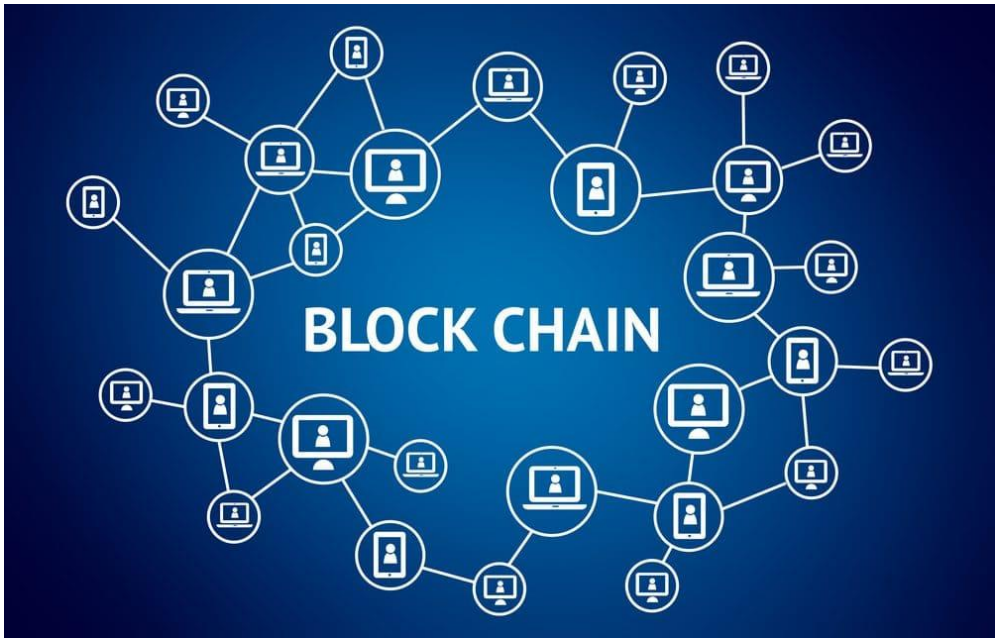
Dear NUS User's,

We Noticed that your NUS Mail Access is out of date and need to be updated for you to remain active.
All out dated account will be suspended finally if you fail to keep your NUS Mail Access updated and active.
We urge you to kindly update your NUS Mail Access and keep it active.
Kindly take a few minute out of your time to keep your NUS Mail Access updated and active to avoid service interruption.

https://www.nus.edu.sg

Regards
National University Of Singapore
Webmal Service Center

Example 3: Document Classification

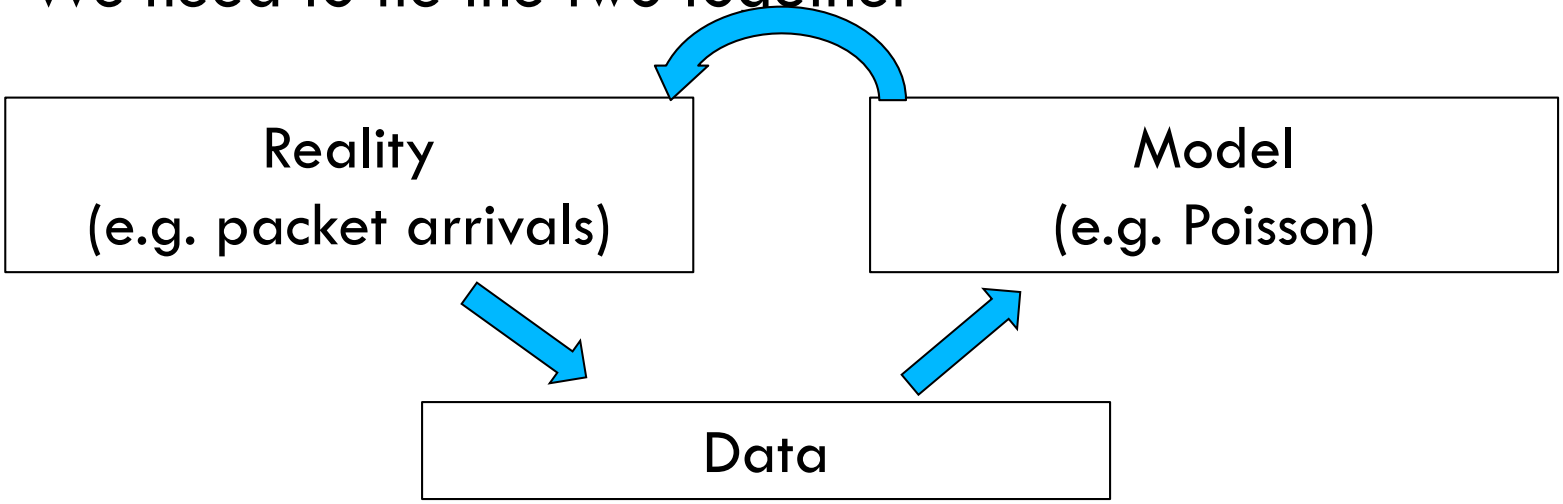


?

- Exchanges
- Wallets and banking
- Payment processing
- Financial products
- Mining
- Gambling
- Service providers
- Information systems
- Security
- Others

Bayesian Statistical Inference

- The world around us has "real" phenomena
- We have looked at models of probabilistic phenomena
- We need to tie the two together



Bayesian Statistical Inference

	Item 1	Item 2	Item 3	Item 4	Item 5
User 1	0	3	0	3	0
User 2	4	0	0	2	0
User 3	0	0	3	0	0
User 4	3	0	4	0	3
User 5	4	3	0	4	0

Netflix Prize: 2009

Bayesian Statistical Inference

Commodity/stock prices

Types of Inference Models

- Model building versus inferring unknown variables
- Consider a system:

$$X = aS + W$$

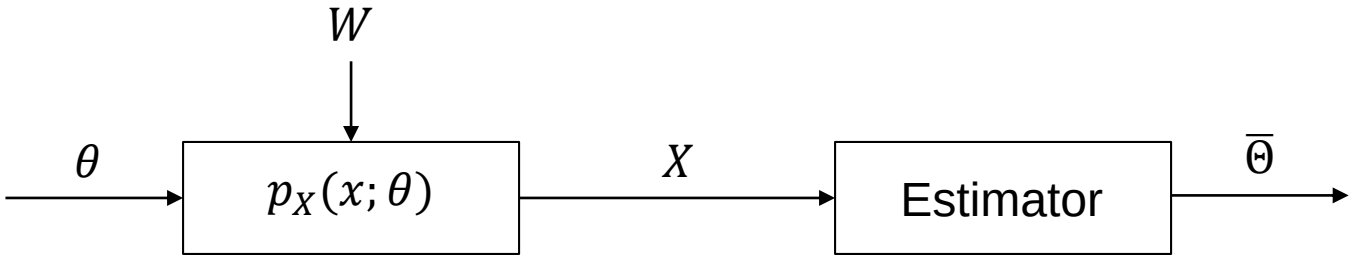
- Model building: We know S, observe X, and try to infer a
- Estimation: We know a, we observe X, and try to infer S
- Types of inference problems:
  - Hypothesis testing: unknown takes a few possible values; minimize the probability of incorrect decision
  - Estimation: minimize the estimation error

Classification, discrete

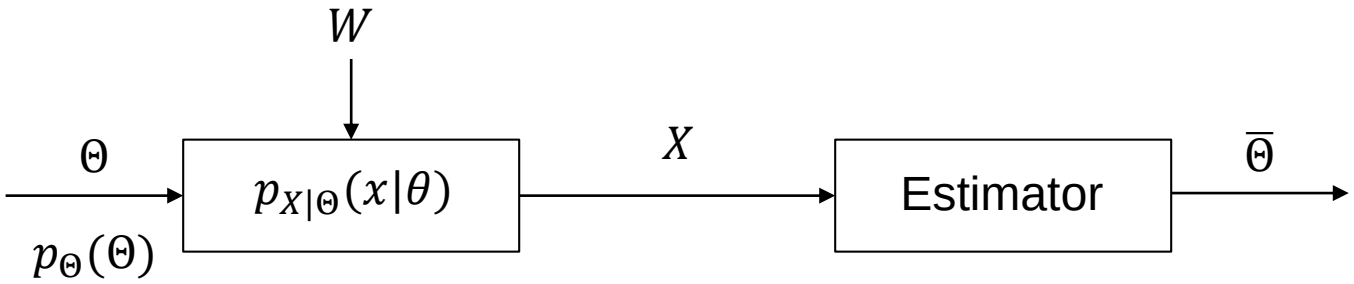
Regression, continuous

Modeling the Unknown Quantity

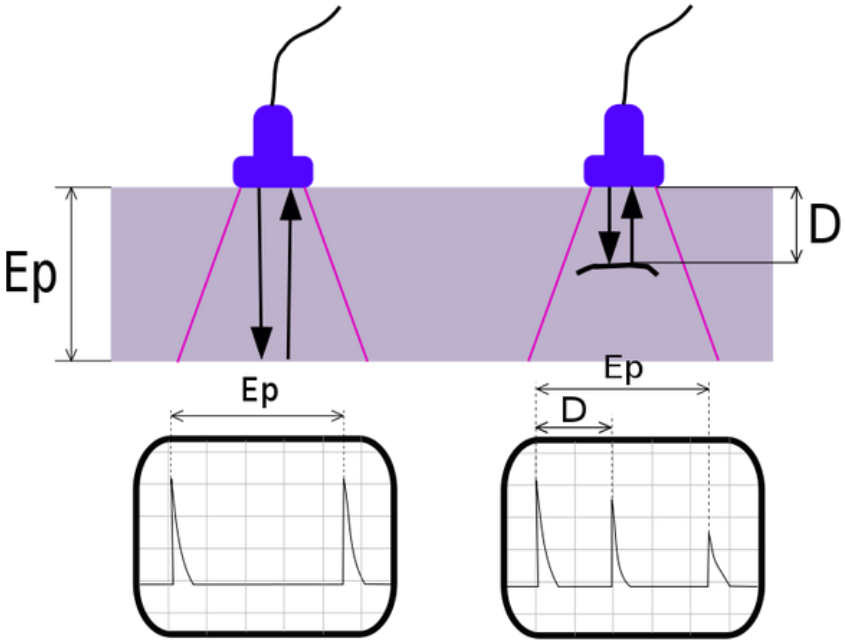
- Philosophical approach 1: The quantity has a definite value (but unknown)



- Philosophical approach 2: We have some initial belief about the unknown quantity (treat as a random variable)

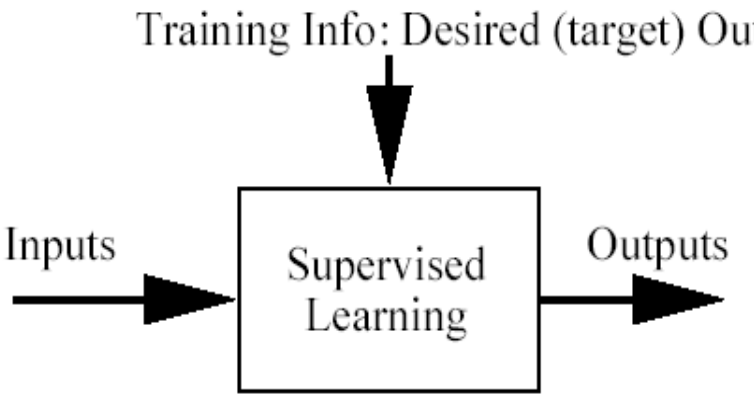


Modeling the Unknown Quantity



Bayesian Inference

- Classification problem:
  - Training data: examples of the form $(d, h(d))$
    - d are the data objects to classify (inputs)
    - $h(d)$ are the correct class info for d, $h(d) \in \{1, \cdots, K\}$
  - Goal: given d_{new}, provide $h(d_{new})$



Bayesian Framework

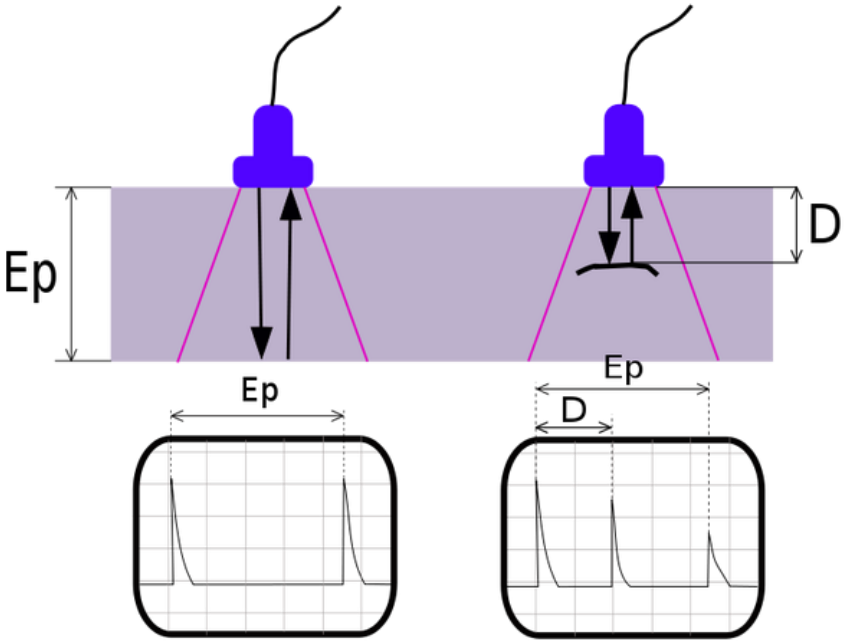
- Allows us to combine observed data and prior knowledge
- Provides practical learning algorithms
- Any kind of objects (e.g. time series, trees, etc.) can be classified, based on a probabilistic model specification

Bayes Rule

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

- Who's who:
  - d: data (observation)
  - h: hypothesis
  - $P(h)$: prior belief (probability of hypothesis h before seeing any data)
  - $P(d|h)$: likelihood (probability of the data if hypothesis h is true)
  - $P(d) = \sum_h P(d|h)P(h)$: data evidence (marginal probability of the data)
  - $P(h|d)$: posterior probability of hypothesis h after seeing the data

Example: Track Monitoring



Example: Track Monitoring

- The detector goes on a mission and one of the tracks tests positive for a crack. It is known that the detector returns a correct positive result in only 98% of the cases and a correct negative result in only 97% of the cases. Furthermore, only 0.8% of the tracks have internal imperfections.
  - What is the probability that this track is damaged?
  - What is the probability that the track is undamaged?
  - What is the diagnosis?

Example: Track Monitoring

- $P(crk) = 0.008$ $\quad P(\neg crk) = 0.992$
- $P(+|crk) = 0.98$ $\quad P(-|crk) = 0.02$
- $P(+|\neg crk) = 0.03$ $\quad P(-|\neg crk) = 0.97$
- $P(+) = P(+|crk)P(crk) + P(+|\neg crk)P(\neg crk) = 0.0376$
- $P(crk|+) = \frac{P(+|crk)P(crk)}{P(+)} = 0.2074$
- $P(\neg crk|+) = \frac{P(+|\neg crk)P(\neg crk)}{P(+)} = 0.7926$
- Diagnosis: not damaged

Naïve Bayes Classifier

Assume target function $f : X \rightarrow V$, where each instance x described by attributes $\langle a_1, a_2 \ldots a_n \rangle$. Most probable value of $f(x)$ is:

$$v_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j | a_1, a_2 \ldots a_n)$$

MAP: maximum a posteriori probability.

ule

of v_j.

ndependence

Naïve Bayes Classifier

Naive_Bayes_Learn($examples$)

For each target value v_j

$\hat{P}(v_j) \leftarrow$ estimate $P(v_j)$

For each attribute value a_i of each attribute a

$\hat{P}(a_i|v_j) \leftarrow$ estimate $P(a_i|v_j)$

Classify_New_Instance(x)

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$$

Example 1

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

- <Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=strong>
- Will Roger play tennis?

Example 1: Solution

features

Day	Outlook	Temperature	Humidity	Wind	PlayTennis (class)
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

some estimates

P(y) = 9/14

P(n) = 5/14

P(sun | y) = 2/9

P(cool | y) = 3/9

P(high | y) = 3/9

P(strong | y) = 3/9

<Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=strong>

P(y)P(sun | y)P(cool | y)P(high | y)P(strong | y) =
(9/14) * (2/9) * (3/9) * (3/9) * (3/9) = 0.005

Example 1: Solution

- <Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=strong>
- Will Roger play tennis?
- Solution: We want to compute

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

  - Two hypotheses:

$$P(y)P(sun|y)P(cool|y)P(high|y)P(strong|y) = 0.005$$
$$P(n)P(sun|n)P(cool|n)P(high|n)P(strong|n) = 0.021$$

- Thus, $v_{NB} = n$
- This is a prediction. If it is sunny, cool, highly humid, and strong wind, it is more likely that we won't play tennis than that we will.

Naïve Bayes Classifier

- Naïve Bayes assumption: Attributes that describe data instances are conditionally independent given the classification hypothesis
  - It is a simplifying assumption and it may be violated in reality
  - In spite of that, it works well in practice
- One of the most practical learning methods
  - Medical Diagnosis
  - Text classification

Practical Limitations

- They typically require initial knowledge of many probabilities
  - If the probabilities are not known in advance, then they have to be estimated based on background knowledge, previously available data, and assumptions about the form of the underlying distributions.
- They have significant computational cost required (linear in the number of candidate hypotheses)

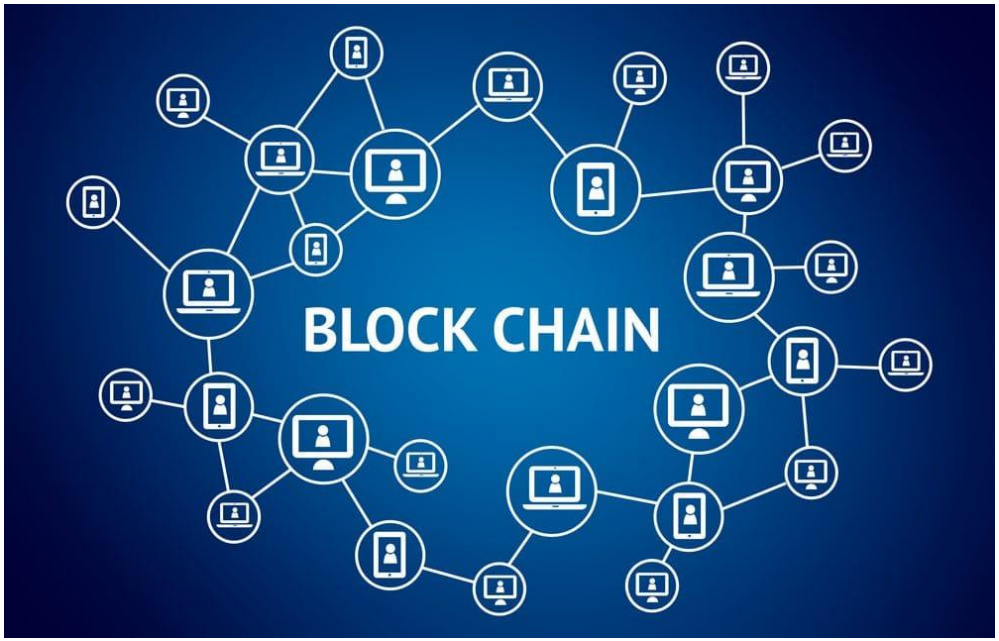
Naïve Bayes Classifier: Subtleties

- What if none of the training instances with target value v_j have attribute value a_i?
- Then: $\hat{P}(a_i|v_j) = 0$ and $\hat{P}(v_j)\prod_i \hat{P}(a_i|v_j) = 0$
- Typical solution:

$$\hat{P}(a_i|v_j) \leftarrow \frac{n_c + mp}{n+m}$$

  - n: number of training examples for which $v = v_j$
  - n_c: number of examples for which $v = v_j$ and $a = a_i$
  - p: prior estimate for $\hat{P}(a_i|v_j)$
  - m: weight given to prior (i.e., number of ``virtual" examples)

Example 3: Document Classification

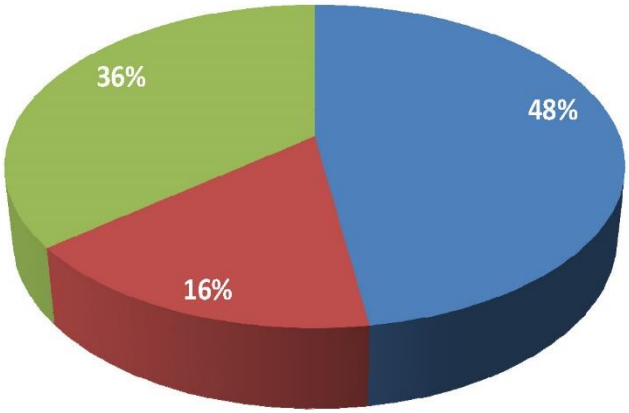


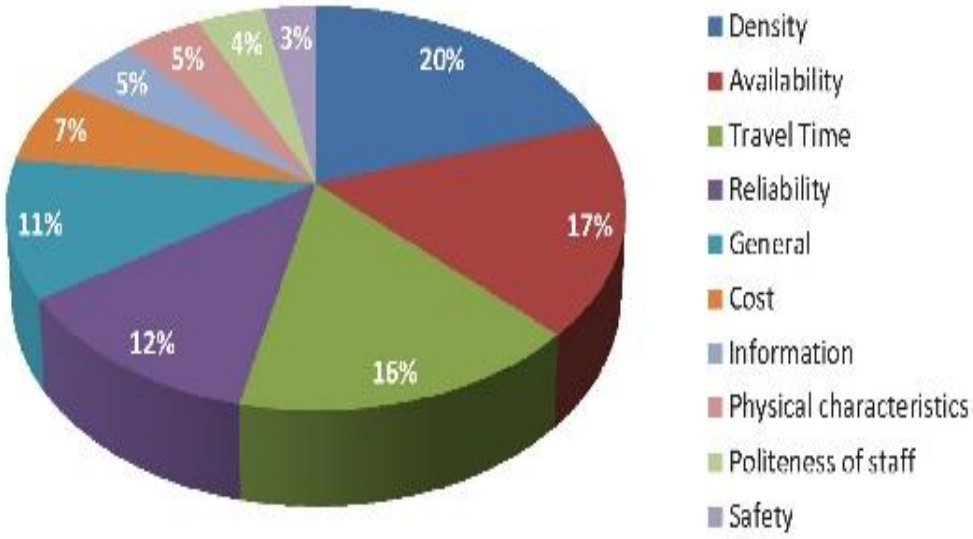
?

- Exchanges
- Wallets and banking
- Payment processing
- Financial products
- Mining
- Gambling
- Service providers
- Information systems
- Security
- Others

Document Classification in IoT?

- @Everton just saw about 30 fans waiting for taxis in lime street station all looking to get to goodison or a pub ahead of the match! #COYB
- Packed at Lime Street whats going on in Liverpool today? (@ Liverpool Lime Street Railway Station (LIV) w/6 others) http://t.co/bISkdpFR





Text Classification

- Application:
  - Learn which news articles/social media posts are of interest
  - Learn to classify web pages by topic
- Naive Bayes is among most effective algorithms for text classification
- What attributes shall we use to represent text documents?

Bag of Words

Y(I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun… It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.)=C

Bag of Words

$Y($ I **love** this movie! It's **sweet**, but with **satirical** humor. The dialogue is **great** and the adventure scenes are **fun**… It manages to be **whimsical** and **romantic** while **laughing** at the conventions of the fairy tale genre. I would **recommend** it to just about anyone. I've seen it **several** times, and I'm always **happy** to see it **again** whenever I have a friend who hasn't seen it yet. $)=C$

Bag of Words


Y(
x love xxxxxxxxxxxxxxxx sweet
xxxxxxx satirical xxxxxxxxxx
xxxxxxxxxxx great xxxxxxx
xxxxxxxxxxxxxxxxxxx fun xxxx
xxxxxxxxxxxxx whimsical xxxx
romantic xxxx laughing
xxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxx recommend xxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xx several xxxxxxxxxxxxxxxxx
xxxxx happy xxxxxxxxx again
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxx
)=C

Text Classification

- Consider a document D, whose class is given by C.
- We classify D as the class which has the highest posterior probability $\mathrm{P}(C|D)$ (from Bayes' Theorem):

$$\mathrm{P}(C|D) = \frac{\mathrm{P}(D|C)\mathrm{P}(C)}{\mathrm{P}(D)}$$

- We will look at two probabilistic models of documents
  - Both represent documents as a bag of words using the Naive Bayes assumption
  - Both represent documents using feature vectors whose components correspond to word types
  - If we have a vocabulary V, containing $|V|$ word types, then the feature vector dimension $d = |V|$

Bag of Words

- Maximum aposterioi probability

$$\underset{c \in C}{\operatorname{argmax}} P(c|D) = \underset{c \in C}{\operatorname{argmax}} P(D|c)P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \cdots, x_n|c)P(c)$$

Represent D by set of features (x_1,...,x_n)

$O(|X|^n \cdot |C|)$ parameters

Can be estimated if a LARGE number of training examples is available

How often does this class occur?

Just count the relative frequencies in a corpus

- Bag of Words assumption: Assume position doesn't matter
- Conditional Independence: Assume the feature probabilities $P(x_i|c)$ are independent given the class c

$$P(x_1, x_2, \cdots, x_n|c) = P(x_1|c)P(x_2|c) \cdots P(x_n|c)$$

Text Classification

- Bernoulli document model: a document is represented by a feature vector with binary elements taking value 1 if the corresponding word is present in the document and 0 if the word is not present.
- Multinomial document model: a document is represented by a feature vector with integer elements whose value is the frequency of that word in the document.
- Example: Consider the vocabulary:

$$V = \{\text{blue, red, dog, cat, biscuit, apple}\}$$

- In this case $d = |V| = 6$

Text Classification

- Consider the (short) document "the blue dog ate a blue biscuit"
- $\mathbf{d}^B$: Bernoulli feature vector for this document

$$\mathbf{d}^B = (1,0,1,0,1,0)^T$$

- $\mathbf{d}^M$: multinomial feature vector

$$\mathbf{d}^M = (2,0,1,0,1,0)^T$$

- To classify a document: estimate the likelihoods of the document given the class, $\mathrm{P}(D|C)$, and the class prior probabilities $\mathrm{P}(C)$
- To estimate the likelihood $\mathrm{P}(D|C)$, we use the Naive Bayes assumption applied to whichever of the two document models we are using.

V = {blue, red, dog, cat, biscuit, apple}

Bernoulli Document Model

- Bernoulli model: document is represented by a binary vector, which represents a point in the space of words.
- If we have a vocabulary V containing a set of $|V|$ words, then the t-th dimension of a document vector corresponds to word w_t in the vocabulary
- $\mathbf{b}_i$: feature vector for the i-th document D_i
- t-th element of $\mathbf{b}_i$, written b_{it}, is either 0 or 1 representing the absence or presence of word w_t in the i-th document.

Bernoulli Document Model

- $\mathrm{P}(w_t|C)$: probability of word w_t occurring in a document of class C
- $1-\mathrm{P}(w_t|C)$: probability of w_t not occurring in a document of class C
- Using naive Bayes assumption:

$$\mathrm{P}(D_i|C) \sim \mathrm{P}(\mathbf{b}_i|C) = \prod_{t=1}^{|V|} \left[b_{it}\mathrm{P}(w_t|C) + (1-b_{it})\left(1-\mathrm{P}(w_t|C)\right) \right]$$

- (Why called Bernoulli: you can think of the document feature vector as being generated by a collection of $|V|$ weighted coin tosses, where the t-th toss has a probability of success (or heads) equal to $\mathrm{P}(w_t|C)$)

Bernoulli Document Model

- Likelihood parameters: $\mathrm{P}(w_t|C)$ and prior probabilities $\mathrm{P}(C)$
- We learn (estimate) these parameters from a training set of documents labelled with class $C = k$
- $n_k(w_t)$: number of documents of class $C = k$ in which w_t is observed
- N_k: the total number of documents of class $C = k$
- Estimates of the parameters of the word likelihoods:

$$\hat{P}(w_t|C = k) = \frac{n_k(w_t)}{N_k}$$

- N: total number of documents in the training set

$$\hat{P}(C = k) = \frac{N_k}{N}$$

Bernoulli Text Classification Model

- Training:
  1. Define the vocabulary V
  2. Count the following in the training set:
     a. N: total number of documents in the training set
     b. N_k: total number of documents of class $C = k$, $k = 1, \cdots, K$
     c. $n_k(w_t)$: number of documents of class $C = k$ containing word w_t, for every class and for each word in the vocabulary
  3. Estimate the likelihoods $\mathrm{P}(w_t|C)$
  4. Estimate the priors $\mathrm{P}(C = k)$

Bernoulli Text Classification Model

- Classification:
- To classify an unlabeled document D_j, we estimate the posterior probability for each class:

$$P(C = k|D_j) = P(C = k|\mathbf{b}_j)$$
$$\propto P(\mathbf{b}_j|C = k)P(C = k)$$
$$\propto P(C = k)\prod_{t=1}^{|V|}\left[b_{jt}\mathrm{P}(w_t|C = k) + (1 - b_{jt})(1 - \mathrm{P}(w_t|C = k))\right]$$

- The selected class:

$$\underset{k}{\operatorname{argmax}}\, P(C = k|D_j)$$

Example

- A set of documents with two classes: Sports (S) or Informatics (I)
- Training set: 11 documents
- Vocabulary of eight words:

$$V = \begin{bmatrix} w_1 = \text{goal} \\ w_2 = \text{tutor} \\ w_3 = \text{variance} \\ w_4 = \text{speed} \\ w_5 = \text{drink} \\ w_6 = \text{defence} \\ w_7 = \text{performance} \\ w_8 = \text{field} \end{bmatrix}$$

- Each document is represented as a 8 dimensional binary vector

Example

- Training data: matrix for each class
- Each row represents a 8-dimensional document vector

$$\mathbf{B}^{S} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$\mathbf{B}^{I} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

- Classify the following:

$\mathbf{b}_1 = (1\ 0\ 0\ 1\ 1\ 1\ 0\ 1)$

$\mathbf{b}_2 = (0\ 1\ 1\ 0\ 1\ 0\ 1\ 0)$

$$V = \begin{bmatrix} w_1 = \text{goal} \\ w_2 = \text{tutor} \\ w_3 = \text{variance} \\ w_4 = \text{speed} \\ w_5 = \text{drink} \\ w_6 = \text{defence} \\ w_7 = \text{performance} \\ w_8 = \text{field} \end{bmatrix}$$

Example

$$\mathbf{B}^{S} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$\mathbf{B}^{I} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

$N = 11$ $N_S = 6$ $N\_I = 5$

$n_s(w_1) = 3$ $n_s(w_2) = 1$ $n_s(w_3) = 2$ $n_s(w_4) = 3$

$n_s(w_5) = 3$ $n_s(w_6) = 4$ $n_s(w_7) = 4$ $n_s(w_8) = 4$

$n_I(w_1) = 1$ $n_I(w_2) = 3$ $n_I(w_3) = 3$ $n_I(w_4) = 1$

$n_I(w_5) = 1$ $n_I(w_6) = 1$ $n_I(w_7) = 3$ $n_I(w_8) = 1$

Example

$$\mathbf{B}^{\mathrm{S}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$\mathbf{B}^{\mathrm{I}} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

$$P(S) = \frac{6}{11} \qquad P(I) = \frac{5}{11}$$

$$P(w_1|S) = \frac{1}{2} \quad P(w_2|S) = \frac{1}{6} \quad P(w_3|S) = \frac{1}{3} \quad P(w_4|S) = \frac{1}{2}$$

$$P(w_5|S) = \frac{1}{2} \quad P(w_6|S) = \frac{2}{3} \quad P(w_7|S) = \frac{2}{3} \quad P(w_8|S) = \frac{2}{3}$$

$$P(w_1|I) = \frac{1}{5} \quad P(w_2|I) = \frac{3}{5} \quad P(w_3|I) = \frac{3}{5} \quad P(w_4|I) = \frac{1}{5}$$

$$P(w_1|I) = \frac{1}{5} \quad P(w_2|I) = \frac{1}{5} \quad P(w_3|I) = \frac{3}{5} \quad P(w_4|I) = \frac{1}{5}$$

Example

$$P(S|\mathbf{b}_1) \propto P(S)\prod_{t=1}^{8}\left[b_{1t}\mathrm{P}(w_t|S)+(1-b_{1t})\left(1-\mathrm{P}(w_t|S)\right)\right]$$

$$=\frac{6}{11}\left(\frac{1}{2}\times\frac{5}{6}\times\frac{2}{3}\times\frac{1}{2}\times\frac{1}{2}\times\frac{2}{3}\times\frac{1}{3}\times\frac{2}{3}\right)=\frac{5}{891}=5.6\times10^{-3}$$

$$P(I|\mathbf{b}_1) \propto P(I)\prod_{t=1}^{8}\left[b_{1t}\mathrm{P}(w_t|I)+(1-b_{1t})\left(1-\mathrm{P}(w_t|I)\right)\right]$$

$$=\frac{5}{11}\left(\frac{1}{5}\times\frac{2}{5}\times\frac{2}{5}\times\frac{1}{5}\times\frac{1}{5}\times\frac{1}{5}\times\frac{2}{5}\times\frac{1}{5}\right)=\frac{8}{859375}=9.3\times10^{-6}$$

Example

$$P(S|\mathbf{b}_2) \propto P(S)\prod_{t=1}^{8}\left[b_{2t}\mathrm{P}(w_t|S)+(1-b_{2t})\left(1-\mathrm{P}(w_t|S)\right)\right]$$

$$=\frac{6}{11}\left(\frac{1}{2}\times\frac{1}{6}\times\frac{1}{3}\times\frac{1}{2}\times\frac{1}{2}\times\frac{1}{3}\times\frac{2}{3}\times\frac{1}{3}\right)=\frac{12}{14256}=8.4\times10^{-4}$$

$$P(I|\mathbf{b}_2) \propto P(I)\prod_{t=1}^{8}\left[b_{2t}\mathrm{P}(w_t|I)+(1-b_{2t})\left(1-\mathrm{P}(w_t|I)\right)\right]$$

$$=\frac{5}{11}\left(\frac{4}{5}\times\frac{3}{5}\times\frac{3}{5}\times\frac{4}{5}\times\frac{1}{5}\times\frac{4}{5}\times\frac{3}{5}\times\frac{4}{5}\right)=\frac{34560}{4296875}=8.0\times10^{-3}$$

Multinomial Document Model

- Multinomial distribution:
  - We have items of d types
  - Proportion of type 1 is p_1, of type 2 is p_2, $\cdots$, of type d is p_d
  - We select n item at random (with replacement)
  - Let x_i denote the number of type i items
  - The probability of observing the vector $\mathbf{x} = (x_1, x_2, \cdots x_d)$:

$$P(\mathbf{x}) = \frac{n!}{x_1!\, x_2! \cdots x_d!} p_1^{x_1} p_2^{x_2} \cdots p_d^{x_d}$$

$$= \frac{n!}{\prod_{t=1}^{d} x_t!} \prod_{t=1}^{d} p_t^{x_t}$$

Example

Multinomial Document Model

- The feature vector in multinomial document captures the frequency of words, not just their presence or absence
- $\mathbf{x}_i$: multinomial model feature vector for document D_i
- x_{it}: number of times word w_t occurs in document D_i
- $n_i = \sum_t x_{it}$: total number of words in document D_i
- $\mathrm{P}(w_t|C)$: probability of word w_t occurring in a document of class C (estimated using the word frequency information from the document feature vectors)

Multinomial Document Model

- The document likelihood:

$$P(D_i|C) \sim P(\mathbf{x}_i|C) = \frac{n!}{\prod_{t=1}^{|V|} x_{it}!} \prod_{t=1}^{|V|} P(w_t|C)^{x_{it}}$$

$$\propto \prod_{t=1}^{|V|} P(w_t|C)^{x_{it}}$$

- Class probabilities: $P(C = k)$

Multinomial Document Model

- Likelihood parameters: $\mathrm{P}(w_t|C)$ and prior probabilities $\mathrm{P}(C)$
- We learn (estimate) these parameters from a training set of documents labelled with class $C = k$
- z_{ik}: indicator variable that equals 1 when D_i has class $C = k$ and equals 0 otherwise
- N: the total number of documents

$$\hat{P}(w_t|C = k) = \frac{\sum_{i=1}^{N} x_{it} z_{ik}}{\sum_{s=1}^{|V|} \sum_{i=1}^{N} x_{is} z_{ik}}$$

Relative frequency of w_t in documents of class $C = k$ with respect to the total number of words in documents of this class

$$\hat{P}(C = k) = \frac{N_k}{N}$$

Multinomial Text Classification Model

- Training:
  1. Define the vocabulary V
  2. Count the following in the training set:
     a. N: total number of documents in the training set
     b. N_k: total number of documents of class $C = k, k = 1, \cdots, K$
     c. x_{it}: the frequency of word w_t in document D_i, computed for every word w_t in V
  3. Estimate the likelihoods $\mathrm{P}(w_t|C)$
  4. Estimate the priors $\mathrm{P}(C = k)$

Multinomial Text Classification Model

- Classification:
- To classify an unlabeled document D_j, we estimate the posterior probability for each class:

$$P(C = k|D_j) = P(C = k|\mathbf{x}_j)$$
$$\propto P(\mathbf{x}_j|C = k)P(C = k)$$
$$\propto P(C = k)\prod_{t=1}^{|V|} \mathrm{P}(w_t|C = k)^{x_{it}}$$

- The selected class:

$$\underset{k}{\operatorname{argmax}}\, P(C = k|D_j)$$

The Zero Probability Problem

- Drawback of relative frequency estimates: zero counts result in estimates of zero probability
- Naive Bayes equation for the likelihood involves taking a product of probabilities
- If any one of the terms of the product is zero, then the whole product is zero
- This means that the probability of the document belonging to the class in question is zero
- Just because a word does not occur in a document class in the training data does not mean that it cannot occur in any document of that class.

The Zero Probability Problem

- Problem: $\hat{P}(w_t|C=k) = \frac{\sum_{i=1}^{N} x_{it} z_{ik}}{\sum_{s=1}^{|V|} \sum_{i=1}^{N} x_{is} z_{ik}}$ underestimates the likelihoods of words that do not occur in the data
- Even if word w is not observed for class $C=k$ in the training set, we would still like $P(w|C=k) > 0$
- Probabilities must sum to 1: if unobserved words have underestimated probabilities, then observed words must have overestimated probabilities.
- "Laplace's law of succession" or "add one smoothing":

$$P_{Lap}(w_t|C=k) = \frac{1 + \sum_{i=1}^{N} x_{it} z_{ik}}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{N} x_{is} z_{ik}}$$

Comparing the two Models

- Underlying model of text and document representation:
  - Bernoulli: word present or not, binary vector
  - Multinomial: count of words, integer vector
- Multiple occurrences of words: Bernoulli: ignored, Multinomial: taken into account
- Behavior with document length: Bernoulli: best for short documents, Multinomial: longer documents are OK
- Behavior with "the":
  - Bernoulli: present in almost every document, $P("\text{the}"|C) = 1$
  - Multinomial: probabilities based on relative frequencies of word occurrence in a class, $P("\text{the}"|C) = 0.05$

Naïve Bayes Summary

- Pros:
  - Very fast and easy-to-implement
  - Well-understood formally and experimentally
- Cons:
  - Seldom gives the very best performance
  - "Probabilities" $P(y|x)$ are not accurate
    - For example, $P(y|x)$ decreases with length of x
    - Probabilities tend to be close to zero or one

Acknowledgements

- A number of the slides in this lecture are based on material from various sources:
  - Dan Jurafsky
  - Tom Mitchell
  - Ata Kaban
  - John Tsitsikilis