Problem Example

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall, short}
  - Country = {G, P}
- Training data: Values of (Hair, Height, Country) collected over population

(B,T,G)  (B,T,P)
(D,T,G)  (B,T,P)
(D,T,G)  (B,T,P)
(D,T,G)  (D,T,P)
(B,T,G)  (D,T,P)
(B,S,G)  (D,S,P)
(B,S,G)  (B,S,P)
(D,S,G)  (D,S,P)
Learn Joint Probabilities

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall, short}
  - Country = {G, P}
- Training data: Values of (Hair, Height, Country) collected over population

### Joint Distribution Table

<table>
<thead>
<tr>
<th>(B,T,G)</th>
<th>(B,T,P)</th>
<th>P(B,S,G)= 2/16</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D,T,G)</td>
<td>(B,T,P)</td>
<td>P(B,T,G)= 2/16</td>
</tr>
<tr>
<td>(D,T,G)</td>
<td>(B,T,P)</td>
<td>P(D,S,G)= 1/16</td>
</tr>
<tr>
<td>(D,T,G)</td>
<td>(D,T,P)</td>
<td>P(D,T,G)= 3/16</td>
</tr>
<tr>
<td>(B,T,G)</td>
<td>(D,T,P)</td>
<td>P(B,S,P)= 1/16</td>
</tr>
<tr>
<td>(B,S,G)</td>
<td>(D,S,P)</td>
<td>P(B,T,P)= 3/16</td>
</tr>
<tr>
<td>(B,S,G)</td>
<td>(B,S,P)</td>
<td>P(D,S,P)= 2/16</td>
</tr>
<tr>
<td>(D,S,G)</td>
<td>(D,S,P)</td>
<td>P(D,T,P)= 2/16</td>
</tr>
</tbody>
</table>

Compute other Joint or Conditional Distributions

\[
P(B,S,G)=\frac{2}{16} \quad P(Hair = B, Height = S | Country = G) = \quad \quad P(Hair = B, Height = S, Country = G) = \quad P(Country = G) = \frac{2}{16} = \frac{4}{16} = \frac{1}{2} \quad P(D,T,P) = \frac{2}{16} \]

\[
P(B,T,G)=\frac{2}{16} \quad P(B,T,P) = \frac{1}{16} \quad P(D,S,G)=\frac{1}{16} \quad P(D,T,G)=\frac{3}{16} \quad P(B,S,P)=\frac{1}{16} \quad P(B,T,P) = \frac{3}{16} \quad P(D,S,P)=\frac{2}{16} \quad P(D,T,P) = \frac{2}{16}
\]
Bayes Classifier Example

• Three variables:
  – Hair = {blond, dark}
  – Height = {tall, short}
  – Country = {G, P}

• Training data: Values of (Hair, Height, Country) collected over population

If I observe a new individual tall with blond hair, what is the most likely country of origin?

(B,T,G) (B,T,P)  
(D,T,G) (B,T,P)  
(D,T,G) (B,T,P)  
(D,T,G) (D,T,P)  
(B,T,G) (D,T,P)  
(B,S,G) (D,S,P)  
(B,S,G) (B,S,P)  
(D,S,G) (D,S,P)

Bayes Classifier

• We want to find the value of $Y$ that is the most probable, given the observations $X_1, \ldots, X_n$

• Find $y$ such that this is maximum:

$$P(Y = y \mid X_1 = x_1, \ldots, X_n = x_n)$$

The maximum is called the Maximum A Posteriori (MAP) estimator
Bayes Classifier

• We want to find the value of $Y$ that is the most probable, given the observations $X_1, \ldots, X_n$

• Find $y$ such that this is maximum:

$$P(Y = y \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{P(X_1 = x_1, \ldots, X_n = x_n \mid Y = y) P(Y = y)}{P(X_1 = x_1, \ldots, X_n = x_n)}$$

Not dependent on $y$
Bayes Classifier

• Classification:
  – Given a new input \((x_1, \ldots, x_n)\), compute the best class:
  \[
  y_{\text{best}} = \arg \max_y \ P(X_1 = x_1, \ldots, X_n = x_n \mid Y = y) \ P(Y = y)
  \]

• Learning:
  – Collect all the observations \((x_1, \ldots, x_n)\) for each class \(y\) and estimate:
  \[
  P(X_1 = x_1, \ldots, X_n = x_n \mid Y = y) = \frac{\text{# observations with } (X_1 = x_1, \ldots, X_n = x_n) \text{ in class } y}{\text{Total Number of observations in class } y}
  \]
  \[
  P(Y = y) = \frac{\text{# observations in class } y}{\text{Total Number of observations}}
  \]

Classifier Example

• Three variables:
  – Hair = \{blond, dark\}
  – Height = \{tall, short\}
  – Country = \{G, P\}

• Training data: Values of (Hair, Height, Country) collected over population

\[
\begin{align*}
(B, T, G) & \quad (B, T, P) & \quad P(B, T|G)P(G) = 2/8 \times 1/2 = 2/16 \\
(D, T, G) & \quad (B, T, P) & \quad P(B, T|P)P(P) = 3/8 \times 1/2 = 3/16 \\
(D, T, G) & \quad (D, T, P) & \\
(B, T, G) & \quad (D, T, P) & \\
(B, S, G) & \quad (D, S, P) & \\
(B, S, G) & \quad (B, S, P) & \\
(D, S, G) & \quad (D, S, P) & \quad \text{Conclusion: Country = P}
\end{align*}
\]

If I observe a new individual tall with blond hair, what is the most likely country of origin?
Classifier Example

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall, short}
  - Country = {G, P}

- Training data: Values of (Hair, Height, Country) collected over population

| (B,T,G)  | (B,T,G)  | (B,T,P)  | P(B,T|G)P(G) = 2/8 x 2/3 = 4/24 |
|----------|----------|----------|---------------------------------|
| (D,T,G)  | (D,T,G)  | (B,T,P)  | P(B,T|P)P(P) = 3/8 x 1/3 = 3/24  |
| (D,T,G)  | (D,T,G)  | (D,T,P)  |                                  |
| (B,T,G)  | (B,T,G)  | (D,T,P)  |                                  |
| (B,S,G)  | (B,S,G)  | (D,S,P)  |                                  |
| (B,S,G)  | (B,S,G)  | (B,S,P)  |                                  |
| (D,S,G)  | (D,S,G)  | (D,S,P)  |                                  |

If I observe a new individual tall with blond hair, what is the most likely country of origin?

Conclusion: Country = G

Naïve Bayes Assumption

To make the problem tractable, we often need to make the following conditional independence assumption:

\[ P(x_1, x_2, \ldots, x_n \mid y) = P(x_1 \mid y)P(x_2 \mid y)\ldots P(x_n \mid y) \]

\[ = \prod_i P(x_i \mid y) \]

which allows us to define the Naïve Bayes Classifier:

\[ y_{NB} = \arg \max_{y \in C} P(y) \prod_i P(x_i \mid y) \]
Naïve Bayes Classifier

- Learning:
  - Collect all the observations \((x_1, ..., x_n)\) for each class \(y\) and estimate:

\[
P(X_i = x_i \mid Y = y) = \frac{\text{Number of observations with } X_i = x_i \text{ in class } y}{\text{Total Number of observations in class } y}
\]

\[
P(Y = y) = \frac{\text{Number of observations in class } y}{\text{Total Number of observations}}
\]

- Classification:

\[
y_{\text{best}} = \arg \max_y P(X_1 = x_1 \mid Y = y) \cdots P(X_n = x_n \mid Y = y) P(Y = y)
\]

How many parameters do we need for the two classifiers: Bayes and Naïve Bayes?
Naïve Bayes Implementation

• Small (but important) implementation detail: If \( n \) is large, we may be taking the product of a large number of small floating-point values. Underflow avoided by taking log.

• Take the max of:

\[
\log P(X_1 = x_1 \mid Y = y) + \ldots + \log P(X_n = x_n \mid Y = y) + \log P(Y = y)
\]

• Instead of:

\[
P(X_1 = x_1 \mid Y = y) \ldots P(X_n = x_n \mid Y = y) P(Y = y)
\]

---

Same Example, the Naïve Bayes Way

• Three variables:
  – Hair = {blond, dark}
  – Height = {tall, short}
  – Country = {G, P}

• Training data: Values of (Hair, Height, Country) collected over population

| (B,T,G) | (B,T,G) | (B,T,P) | \( P(B,T|G)P(G) \approx P(B|G)P(T|G)P(G) \) |
|---------|---------|---------|------------------------------------------|
| (D,T,G) | (D,T,G) | (B,T,P) | 8/16 x 10/16 x 2/3 \( \approx 160/768 = 40/192 \) |
| (D,T,G) | (D,T,G) | (B,T,P) | \( P(B,T|P)P(P) \approx 4/8 x 5/8 x 1/3 = 20/192 \) |
| (B,T,G) | (B,T,G) | (D,T,P) | Conclusion: Country = G |
| (B,S,G) | (B,S,G) | (D,S,P) | |
| (B,S,G) | (B,S,G) | (B,S,P) | |
| (D,S,G) | (D,S,G) | (D,S,P) | |
Same Example, the Naïve Bayes Way

- Three variables:
  - Hair = {blond, dark}
  - Height = {tall, short}
  - Country = {G, P}
- Training data: Values of (Hair, Height, Country) collected over population

The variables are not independent so it is only an approximation.

The values are of course different, but the conclusion remains the same:
- 0.17 vs. 0.2 for Country = G
- 0.125 vs. 0.1 for Country = P

| (B,T,G) | (B,G) | (T,G) | P(B|G)P(T|G)P(G) |
|--------|-------|-------|------------------|
| (D,T,G) | (D,G) | (T,G) | P(B|G)P(T|G)P(G) |
| (D,T,G) | (D,G) | (P,G) | P(B|G)P(T|G)P(G) |

Conclusion: Country = G

Naïve Bayes Classifier

Yet another classifier.

When to use?
- Moderate or large training set available
- Attributes that describe instances are conditionally independent given class

Successful applications:
- Diagnosis
- Classifying text documents
Naïve Bayes: Subtleties

Conditional independence assumption is often violated

\[ P(x_1, x_2, \ldots, x_n \mid y) = \prod_i P(x_i \mid y) \]

... but it works surprisingly well anyway.

A plausible reason is that to make correct predictions,

- Don’t need the probabilities to be estimated correctly
- Only need the posterior of the correct class to be largest among the class posteriors

\[
\arg\max_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_i \mid v_j) = \arg\max_{v_j \in V} P(v_j)P(a_1 \ldots, a_n \mid v_j)
\]

Posterior are often unrealistically close to 0 or 1

Naïve Bayes: Subtleties

What if none of the training instances with target values \( v_j \) have attribute value \( a_i \)? Then

\[
\hat{P}(a_i \mid v_j) = 0, \text{ and...} \\
\hat{P}(v_j) \prod_i \hat{P}(a_i \mid v_j) = 0
\]

Add one smoothing:

\[
p(a_i \mid v_j) = \frac{n_{a_i} + 1}{n + |M|}
\]

\( M \): # of possible values of \( a_i \)
Naïve Bayes: Subtleties

General solution is Bayesian estimate (smoothing):

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

Where: 
- \( n \) is number of training examples for which \( v = v_j \)
- \( n_c \) is number of examples for which \( v = v_j \) and \( a = a_i \)
- \( p \) is prior estimate for \( P(a_i | v_j) \)
- \( m \) is weight given to prior (i.e., number of 'virtual' examples)

Naïve Bayes in Text Classification

Classes can be:
- topics (politics, business, entertainment, sports, etc.)
- spam vs. non-spam email
- positive vs. negative opinion
- Many others

Naïve Bayes is among the most effective algorithms

What attributes shall we use to represent text documents?
Text Classification

Represent each document by vector of words

Naïve Bayes conditional independence assumption:

\[ P(doc \mid v_j) = \prod_{i=1}^{len(doc)} P(a_i = w_k \mid v_j) \]

Multinomial distribution

One more assumption, position doesn’t matter

\[ P(a_i = w_k \mid v_j) = P(a_m = w_k \mid v_j), \forall i, m \]

Bag-of-word model, multinomial naïve Bayes classifier

---

Multinomial Naïve Bayes: Learning

- From training corpus, extract Vocabulary
- Calculate \( P(c_j) \)
  - For each \( c_j \) in \( C \) do
    - \( docs_j \leftarrow \) all docs with class \( = c_j \)
    - \( P(c_j) \leftarrow \frac{|docs_j|}{\text{total # documents}} \)
- Calculate \( P(w_k \mid c_j) \)
  - \( Text_j \leftarrow \) single doc containing all \( docs_j \)
  - For each word \( w_k \) in Vocabulary
    - \( n_k \leftarrow \) # of occurrences of \( w_k \) in \( Text_j \)
    - \( P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|} \)

\( \alpha = 1 \) add-one smoothing
Multinomial Naïve Bayes: Testing

- Return $C_{NB}$ where

$$c_{NB} = \operatorname*{argmax}_{c_j \in C} P(c_j) \prod_{i \in \text{positions}} P(x_i | c_j)$$

Generative vs. Discriminative Models

Given training examples $(x_1, y_1), \ldots, (x_n, y_n)$,

<table>
<thead>
<tr>
<th><strong>Discriminative Models</strong></th>
<th><strong>Generative Models</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Select hypothesis space $H$ to consider</td>
<td>Select set of distributions to consider for modeling $P(X,Y)$</td>
</tr>
<tr>
<td>Find $h$ from $H$ with lowest training error</td>
<td>Find distribution that best matches $P(X,Y)$ on training data</td>
</tr>
<tr>
<td>Argument: low training error leads to low prediction error</td>
<td>Argument: If match is close enough, we can use Bayes decision rule</td>
</tr>
<tr>
<td>Examples: decision trees, perceptrons, SVMs</td>
<td>Examples: naïve Bayes, HMMs</td>
</tr>
</tbody>
</table>
Generative Model for Multinomial Naïve Bayes

\[ \hat{P}(w \mid c) = \frac{\text{count}(w, c) + 1}{\text{count}(c) + |V|} \]

Text Classification Example

<table>
<thead>
<tr>
<th>Doc</th>
<th>Words</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training 1</td>
<td>Chinese Beijing Chinese</td>
<td>c</td>
</tr>
<tr>
<td>2</td>
<td>Chinese Chinese Shanghai</td>
<td>c</td>
</tr>
<tr>
<td>3</td>
<td>Chinese Macao</td>
<td>c</td>
</tr>
<tr>
<td>4</td>
<td>Tokyo Japan Chinese</td>
<td>j</td>
</tr>
<tr>
<td>Test 5</td>
<td>Chinese Chinese Chinese Tokyo Japan</td>
<td>?</td>
</tr>
</tbody>
</table>

Priors:
- \[ P(c) = \frac{3}{4} \]
- \[ P(j) = \frac{1}{4} \]

Choosing a class:
- \[ P(c \mid dS) = \frac{3}{4} \times \frac{(3/7)^3}{1/4} \times 1/14 = 0.0003 \]

Conditional Probabilities:
- \[ P(\text{Chinese} \mid c) = \frac{(5+1)}{(8+6)} = 6/14 = 3/7 \]
- \[ P(\text{Tokyo} \mid c) = \frac{(0+1)}{(8+6)} = 1/14 \]
- \[ P(\text{Japan} \mid c) = \frac{(0+1)}{(8+6)} = 1/14 \]
- \[ P(\text{Chinese} \mid j) = \frac{(1+1)}{(3+6)} = 2/9 \]
- \[ P(\text{Tokyo} \mid j) = \frac{(1+1)}{(3+6)} = 2/9 \]
- \[ P(\text{Japan} \mid j) = \frac{(1+1)}{(3+6)} = 2/9 \]

Slide from Dan Jurafsky
So Far

Bayes classifier and Naïve Bayes classifier
Applications

Next: Bayes rule in choosing hypothesis

Hypothesis Selection: An Example

I have three identical boxes labeled H1, H2 and H3.
Into H1 I place 1 black bead, 3 white beads.
Into H2 I place 2 black beads, 2 white beads.
Into H3 I place 4 black beads, no white beads.

I draw a box at random.
I remove a bead at random from that box, take note of its color, and put it back into the box. I repeat this process to generate a sequence of colors.
Which box is most likely to have yielded this color sequence?
Bayesian Solution

A nice way to look at this:

- H1, H2 and H3 were my prior models of the world.
- The fact that \( P(H1) = 1/3, P(H2) = 1/3, P(H3) = 1/3 \) was my prior distribution.
- The color of the bead was a piece of evidence about the true model of the world.
- The use of bayes’ rule was probabilistic inference, giving me a posterior distribution on possible worlds.
- Learning is prior + evidence ---> posterior
- A piece of evidence decreases my ignorance about the world.

Bayesian Methods for Hypothesis Selection

**Goal:** find the best hypothesis from some space \( H \) of hypotheses, given the observed data \( D \).

The learner considers a set of candidate hypotheses \( H \) (models), and attempts to find the most probable one \( h \in H \), given the observed data.

Such maximally probable hypothesis is called maximum a posteriori hypothesis (MAP); Bayes theorem is used to compute it:

\[
\begin{align*}
    h_{MAP} &= \arg\max_{h \in H} P(h \mid D) = \arg\max_{h \in H} P(D \mid h) \frac{P(h)}{P(D)} \\
    &= \arg\max_{h \in H} P(D \mid h) P(h)
\end{align*}
\]
Notations

\( P(h) \) - the prior probability of a hypothesis \( h \)
Reflects background knowledge; before data is observed.
If no information - uniform distribution

\( P(D) \) - The probability that this sample of the Data is observed. (No knowledge of the hypothesis)

\( P(D|h) \) - The probability of observing the sample \( D \), given that the hypothesis \( h \) holds

\( P(h|D) \) - The posterior probability of \( h \). The probability \( h \) holds, given that \( D \) has been observed.

Bayes Theorem

\[
P(h|D) = \frac{P(D|h)P(h)}{P(D)}
\]

\( P(h|D) \) increases with \( P(h) \) and with \( P(D|h) \)

\( P(h|D) \) decreases with \( P(D) \)
Learning Scenario

\[ h_{\text{MAP}} = \arg\max_{h \in H} P(h \mid D) = \arg\max_{h \in H} P(D \mid h)P(h) \]

We may assume that a priori, hypotheses are equally probable

\[ P(h_i) = P(h_j), \forall h_i, h_j \in H \]

We get the Maximum Likelihood hypothesis:

\[ h_{\text{ML}} = \arg\max_{h \in H} P(D \mid h) \]

This way we just look for the hypothesis that best explains the data

---

Example

\[ h_{\text{MAP}} = \arg\max_{h \in H} P(h \mid D) = \arg\max_{h \in H} P(D \mid h)P(h) \]

A given coin is either fair or has a 60% bias in favor of Tail. Decide whether the coin is fair or biased

Two hypotheses: \( h_1 \text{ fair} \): \( P(T)=0.5 \); \( h_2 \text{ biased} \): \( P(T)=0.6 \)

Prior: \( P(h) \):
\( P(h_1)=0.75 \) \( P(h_2)=0.25 \) This is given.

Now we need Data. 1st Experiment: coin toss is T.

\[ P(D|h): \quad P(D|h_1)=0.5 \ ; \ P(D|h_2)=0.6 \]

\[ P(D): \quad P(D)=P(D|h_1)P(h_1)+P(D|h_2)P(h_2) \]
\[ = 0.5 \times 0.75 + 0.6 \times 0.25 = 0.525 \]

\[ P(h|D): \]
\( P(h_1|D)=P(D|h_1)P(h_1)/P(D)=0.5 \times 0.75/0.525=0.714 \)
\( P(h_2|D)=P(D|h_2)P(h_2)/P(D)=0.6 \times 0.25/0.525=0.286 \)
A given coin is either fair or has a 60% bias in favor of Tail. Decide whether the coin is fair or biased.

Two hypotheses: $h_1$ fair: $P(T)=0.5$; $h_2$ biased: $P(T)=0.6$

Prior: $P(h)$: $P(h_1)=0.75$  $P(h_2)=0.25$  This is given.

After 1st coin toss is $T$ we still think that the coin is more likely to be fair.

If we were to use Maximum Likelihood approach (i.e., assume equal priors) we would think otherwise. The data supports the biased coin better.

Try: 100 coin tosses; 70 tails.
Now you will believe that the coin is biased.

Example (2)

$h_{MAP} = \arg\max_{h \in H} P(h \mid D) = \arg\max_{h \in H} P(D \mid h)P(h)$

A given coin is either fair or has a 60% bias in favor of Tail. Decide whether the coin is fair or biased.

Two hypotheses: $h_1$ fair: $P(T)=0.5$; $h_2$ biased: $P(T)=0.6$

Prior: $P(h)$: $P(h_1)=0.75$  $P(h_2)=0.25$  This is given.

Case of 100 coin tosses; 70 tails.

$$P(D) = P(D \mid h_1)P(h_1) + P(D \mid h_2)P(h_2)$$

$$= 0.5^{100} \cdot 0.75 + 0.6^{70} \cdot 0.4^{30} \cdot 0.25 = 7.9 \cdot 10^{-31} \cdot 0.75 + 3.4 \cdot 10^{-28} \cdot 0.25$$

$$P(h_1 \mid D) = \frac{P(D \mid h_1)P(h_1)}{P(D)} <\{ P(D \mid h_2)P(h_2) \}/P(D) = P(h_2 \mid D)$$

0.0057 0.9943
Summary

• Basic probability concepts
• Bayes rule
• What are joint distributions
• Inference using joint distributions
• Learning joint distributions from data
• Independence
• Bayes classifiers
• Naïve Bayes approach
• Selecting hypothesis using Bayesian methods