Inf2b Learning and Data
Lecture 5: Introduction to statistical pattern recognition

Hiroshi Shimodaira
(Credit: Iain Murray and Steve Renals)

Centre for Speech Technology Research (CSTR)
School of Informatics
University of Edinburgh

Jan-Mar 2015
1. Probability (review)

2. What is Bayes’ theorem for?

3. Bayes decision rule

4. More about probability – pitfalls
Motivation for probability

In some applications we need to:

- Communicate uncertainty
- Use prior knowledge
- Deal with missing data
  (we cannot easily measure similarity)
Warming up

- Throwing two dices
  - Probability of \{1, 1\}?
  - Probability of \{2, 5\}?

- Drawing two cards from a deck of cards
  - Probability of \{Club, Spade}\?
  - Probability of \{Club, Club}\?
Probability that a student in Informatics has eyeglasses?

Probability that you live more than 90 years?

When a real dice is thrown, is the probability of getting \{1\} \frac{1}{6}? 

Theoretical probability vs. Empirical probability

aka:
relative frequency
experimental probability
Rules of Probability

<table>
<thead>
<tr>
<th>Random variables</th>
<th>Events/values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>${x_1, x_2, \ldots, x_L}$</td>
</tr>
<tr>
<td>$Y$</td>
<td>${y_1, y_2, \ldots, y_M}$</td>
</tr>
</tbody>
</table>

**Product Rule:**

\[
P(Y = y_j, X = x_i) = P(Y = y_j \mid X = x_i) P(X = x_i)
\]

\[
= P(X = x_i \mid Y = y_j) P(Y = y_j)
\]

Abbreviation:

\[
P(Y, X) = P(Y \mid X) P(X)
\]

\[
= P(X \mid Y) P(Y)
\]

$X$ and $Y$ are independent iff:

\[
P(X \mid Y) = P(X), \quad P(Y \mid X) = P(Y)
\]

\[
P(X, Y) = P(X) P(Y)
\]
Rules of Probability (cont.)

Sum Rule:

\[ P(X = x_i) = \sum_{j=1}^{M} P(X = x_i, Y = y_j) \]

Abbreviation:

\[ P(X) = \sum_{Y} P(X, Y) \]

RHS: Marginalisation of the joint probability over \( Y \).

LHS: Marginal probability of \( X \).

Application:

\[ P(X) = \sum_{Y} P(X | Y) P(Y) \]
1 Probability (review)

2 What is Bayes’ theorem for?

3 Bayes decision rule

4 More about probability – pitfalls
Example: determining the sex of fish

Histograms of fish lengths \((N_F = N_M = 100)\)

Lengths of male fish

Lengths of female fish
Example: determining the sex of fish

Relative frequencies of fish length

Lengths of male fish

Lengths of female fish
Example: determining the sex of fish

Possible decision boundary

Lengths of male fish

Lengths of female fish

Rel. Freq.
Length / cm

Ref. Freq.
Length / cm
Fish questions

- How to classify 4 cm, or 19 cm fish?
- How to classify 10 cm fish?

Lengths of male fish

Lengths of female fish
Fish questions

Relative frequency of male fish length: \( P(x \mid M) \)
Relative frequency of female fish length: \( P(x \mid F) \)

Given a fish length, \( x \), is it sensible to decide as follows?

If \( P(x \mid M) > P(x \mid F) \) ⇒ male fish
If \( P(x \mid M) < P(x \mid F) \) ⇒ female fish
Fish questions (cont.)

How to obtain $P(M \mid x)$ and $P(F \mid x)$?

- The product rule:
  \[
  P(M, x) = P(M \mid x) P(x) = P(x \mid M) P(M)
  \]

- Posterior probabilities:
  \[
  P(M \mid x) = \frac{P(x \mid M) P(M)}{P(x)} \propto P(x \mid M) P(M)
  \]
  \[
  P(F \mid x) = \frac{P(x \mid F) P(F)}{P(x)} \propto P(x \mid F) P(F)
  \]
Bayes’ Theorem

\[ P(H \mid E) = \frac{P(E \mid H) P(H)}{P(E)} \]

Thomas Bayes (?) (1702? – 1761)

http://www.york.ac.uk/depts/maths/histstat/bayespic.htm

c.f. Bayesian inference, Bayesian
LII. *An Essay towards solving a Problem in the Doctrine of Chances.* By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S.

Dear Sir,

Read Dec. 23, 1763. I now send you an essay which I have found among the papers of our deceased friend Mr. Bayes, and which, in my opinion, has great merit, and well deserves to be preserved. Experimental philosophy, you will find, is nearly interested in the subject of it; and on this account there seems to be particular reason for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illustrious Society, and was much esteemed by many in it as a very able mathematician. In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circumstances, it has happened a certain number of times, and failed a certain other number of times.
Non-examinable!

Bayes’ paper:
http://www.jstor.org/stable/105741
http://dx.doi.org/10.1093/biomet/45.3-4.296 (re-typeset)

Cox’s paper:
http://dx.doi.org/10.1119/1.1990764
http://dx.doi.org/10.1016/S0888-613X(03)00051-3 modern commentary

MacKay textbook, amongst many others
1. Probability (review)

2. What is Bayes’ theorem for?

3. Bayes decision rule

4. More about probability – pitfalls
Bayes decision rule

Class $C = \{c_1, \ldots, c_K\}$; input features $X = \mathbf{x}$

**Choose the most probable class:** (maximum posterior class)

$$c^* = \arg \max_{c_k} P(c_k | \mathbf{x}) = \arg \max_{c_k} P(\mathbf{x} | c_k) P(c_k)$$

where

$$P(c_k | \mathbf{x}) = \frac{P(\mathbf{x} | c_k) P(c_k)}{P(\mathbf{x})} = \frac{P(\mathbf{x} | c_k) P(c_k)}{\sum_{j=1}^K P(\mathbf{x} | c_j) P(c_j)}$$

⇒

- It is known this decision rule gives minimum error rate. (We will discuss this in Lecture 10)
- Also called
  - Minimum error (misclassification) rate classification (PRML C. M. Bishop (2006) Section 1.5)
  - Maximum posterior probability (MAP) decision rule
Inferring labels for $x = 11$

<table>
<thead>
<tr>
<th>Classify $x$ as</th>
<th>Posterior probs</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>$P(M \mid x) &gt; P(F \mid x)$ ⇔ $\frac{P(M \mid x)}{P(F \mid x)} &gt; 1$</td>
<td></td>
</tr>
<tr>
<td>$F$</td>
<td>$P(M \mid x) &lt; P(F \mid x)$ ⇔ $\frac{P(M \mid x)}{P(F \mid x)} &lt; 1$</td>
<td></td>
</tr>
</tbody>
</table>

- **Equal prior probabilities:** classify it as male:

  
  
  $$P(M \mid x = 11) = \frac{P(x = 11 \mid M) P(M)}{P(x = 11)}$$

  
  $$= \frac{P(x = 11 \mid M) P(M)}{P(x = 11 \mid M) P(M) + P(x = 11 \mid F) P(F)}$$

  
  $$= \frac{0.14 \cdot 0.5}{0.14 \cdot 0.5 + 0.10 \cdot 0.5} = \frac{0.14}{0.24} = 0.583$$

  
  $$P(F \mid x = 11) = \frac{P(x = 11 \mid F) P(F)}{P(x = 11 \mid M) P(M) + P(x = 11 \mid F) P(F)}$$

  
  $$= \frac{0.10 \cdot 0.5}{0.14 \cdot 0.5 + 0.10 \cdot 0.5} = \frac{0.10}{0.24} = 0.416$$
Inferring labels for $x = 11$ (cont.)

- **Equal prior probabilities:** classify it as male:

  \[
  \frac{P(M | x = 11)}{P(F | x = 11)} = \frac{P(x = 11 | M) P(M)}{P(x = 11 | F) P(F)} = \frac{0.14 \cdot 0.5}{0.10 \cdot 0.5} = 1.4
  \]

- **Twice as many females as males:** (i.e., $P(M) = 1/3$, $P(F) = 2/3$)

  \[
  \frac{P(M | x = 11)}{P(F | x = 11)} = \frac{P(x = 11 | M) P(M)}{P(x = 11 | F) P(F)} = \frac{0.14 \cdot 1/3}{0.10 \cdot 2/3} = 0.7
  \]

  Classify it as female
Likelihood vs posterior probability

\[ P(c_k|x) = \frac{P(x|c_k)P(c_k)}{P(x)} \]

\[ P(M) : P(F) = 1 : 1 \]
Likelihoods vs posterior probabilities (cont.)

\[ P(c_k|x) = \frac{P(x|c_k)P(c_k)}{P(x)} \]

\[ P(M) : P(F) = 1 : 4 \]
Some more questions

• Assume $P(M) = P(F) = 0.5$

1. What is the value of $P(M \mid X = 4)$?
2. What is the value of $P(F \mid X = 18)$?
3. You observe data point $x = 22$. To which class should it be assigned?

• Discuss how classification performance can be improved.
  • What if we increase the number of histogram bins?
  • What if we increase the number of samples?
  • What if we measure not only fish length but also weight? (How can we estimate probabilities?)

• It seems that we can estimate $P(C \mid x)$ directly from data, right?
1. Probability (review)

2. What is Bayes’ theorem for?

3. Bayes decision rule

4. More about probability – pitfalls
More about probability

- Conditional probability of three variables

\[
P(X, Y \mid Z) = \frac{P(Y, Z \mid X) P(X)}{P(Z)}
\]

\[
P(X \mid Y, Z) = \frac{P(Z \mid Y, X) P(X \mid Y)}{P(Z \mid Y)}
\]

- Chain rule

\[
P(X_1, X_2, \ldots, X_N) = P(X_1) P(X_2 \mid X_1) P(X_3 \mid X_1, X_2) \cdots \cdots P(X_N \mid X_1, \ldots, X_{N-1})
\]

Prove!
Independence vs Pearson’s correlation coefficient $\rho = 0$

If $X$ and $Y$ are independent, $\rho_{XY} = 0$. The converse is not true.

See https://en.wikipedia.org/wiki/Correlation_and_dependence

E.g. $(X, Y) = (-1, 0), (0, -1), (0, 1), (1, 0)$, each of which occurs with a probability of $\frac{1}{4}$.

\[
\begin{align*}
P(X = -1) P(Y = 0) &= 1/4 \cdot 1/2 = 1/8 \\
P(X = 0) P(Y = -1) &= 1/2 \cdot 1/4 = 1/8 \\
&\quad P(X = 0) P(Y = 1) = 1/2 \cdot 1/4 = 1/8 \\
&\quad P(X = 1) P(Y = 0) = 1/4 \cdot 1/2 = 1/8
\end{align*}
\]

$\rho_{XY} = 0$, but $P(X, Y) \neq P(X) P(Y)$

i.e., not independent
The Monty Hall problem

Suppose you’re on a game show, and you’re given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what’s behind the doors, opens another door, say No. 3, which has a goat. He then says to you, ”Do you want to pick door No. 2?” Is it to your advantage to switch your choice?

For details, see:  
http://en.wikipedia.org/wiki/Monty_Hall_problem  
Summary

- Bayes’ theorem for statistical pattern classification
- Posterior is proportional to prior times likelihood
- \( P(x) \) can be obtained with marginalisation of \( P(x|C)P(C) \)
- Bayes decision rule achieves minimum error rate classification
- Discuss possible difficulties of applying the Bayes’ decision rule to real problems