**Today’s Schedule**

1. Probability (review)
2. What is Bayes’ theorem for?
3. Bayes decision rule
4. More about probability
5. Optimisation problems

**Motivation for probability**

In some applications we need to:
- Communicate uncertainty
- Use prior knowledge
- Deal with missing data

(we cannot easily measure similarity)

**Introduction to statistical pattern recognition and Optimisation**

Hiroshi Shimodaira

(Credit: Iain Murray and Steve Renals)

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

http://www.inf.ed.ac.uk/teaching/courses/inf2b/
https://psaszza.com/ed.ac.uk/spring2018/inf2b0909learning

Office hours: Wednesdays at 14:00-15:00 in IF-3.04

Jan-Mar 2018

---

**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)?

**Example: determining the sex of fish**

Relative frequencies of fish length

Lengths of male fish

Lengths of female fish

(NB: different example from the one in Note 5.)

---

**Rules of Probability**

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

**Product Rule:**

$P(Y = y_j | X = x_i) = P(Y = y_j | X = x_i) P(X = x_i)$

$= P(X = x_i | Y = y_j) P(Y = y_j)$

**Abbreviation:**

$P(Y, X) = P(Y | X) P(X)$

$= P(X | Y) P(Y)$

$X$ and $Y$ are independent if:

$P(X | Y) = P(X)$

$P(Y | X) = P(Y)$

---

**Example: determining the sex of fish**

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

Lengths of male fish

Relative frequencies of fish length

Lengths of male fish

Lengths of female fish

(NB: different example from the one in Note 5.)
Example: determining the sex of fish

Possible decision boundary

Rel. Freq. of male fish length: $P(x | M)$

Rel. Freq. of female fish length: $P(x | F)$

Given a fish length, $x$, is it sensible to decide as follows?

- If $P(x | M) > P(x | F)$ ⇒ male fish
- If $P(x | M) < P(x | F)$ ⇒ female fish

Fish questions (cont.)

How to obtain $P(Y | x)$? (where $Y = \{F, M\}$)

- The product rule:
  $$P(Y, x) = P(x | Y) P(Y)$$

- Posterior probabilities:
  $$P(Y | x) = \frac{P(x | Y) P(Y)}{P(x)}$$

Bayes’ Theorem

$$P(H | E) = \frac{P(E | H) P(H)}{P(E)}$$

Thomas Bayes (?) (1702 – 1761)

http://www.york.ac.uk/depgs/mathhist/bayespic.htm

c.f. Bayesian inference, Bayesian

Fish questions

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

Bayes’ decision rule

Class $C = \{1, \ldots, K\}$, $C_k$ to denote $C = k$; input features $X = x$

Choose the most probable class: (maximum posterior class)

$$k_{\text{max}} = \arg \max_{k \in C} P(C_k | x) = \arg \max_{k \in C} P(x | C_k) P(C_k)$$

where

$$P(C_k | x) = \frac{P(x | C_k) P(C_k)}{P(x)}$$

$\Rightarrow$ It is known this decision rule gives minimum error rate.

(We will discuss this in Lecture 10)

Also called

- Minimum error (misclassification) rate classification
- Maximum posterior probability (MAP) decision rule

Inferring labels for $x = 11$

- Equal prior probabilities:
  $$P(M | x = 11) = \frac{P(x = 11 | M) P(M)}{P(x = 11)}$$
  $$= \frac{P(x = 11 | M) P(M)}{P(x = 11 | M) P(M) + P(x = 11 | F) P(F)}$$
  $$= \frac{0.14 \cdot 0.5 + 0.10 \cdot 0.5}{0.14 \cdot 0.5 + 0.10 \cdot 0.5} = 0.583$$

$$P(F | x = 11) = \frac{P(x = 11 | F) P(F)}{P(x = 11)}$$
  $$= \frac{0.14 \cdot 0.5 + 0.10 \cdot 0.5}{0.14 \cdot 0.5 + 0.10 \cdot 0.5} = 0.416$$

$\Rightarrow$ classify it as male

NB: For classification, no need to calculate $P(x = 11)$. 

Non-examinable!

Bayes’ paper:
http://www.jstor.org/stable/100741
http://dx.doi.org/10.1093/biomstat/45.3.4.296 [re-print]

Cox’s paper:
http://dx.doi.org/10.1119/1.1900764
http://dx.doi.org/10.1063/30888-613X(03)00051-3 [online commentary]

MacKay textbook, amongst many others
Introduction to statistical pattern recognition and Optimisation

What is the value of $P(x|C)$? For UG4 projects, find the optimal allocation of supervisors and students under given constraints (e.g. no supervisors can take more than five students.)

Likelihood vs posterior probability

- Equal prior probabilities:
  
  $$P(C|x) = \frac{P(x|C)P(C)}{P(x)} = \frac{\sum_{i=1}^{n} P(x_i|C)P(C)}{P(x)}$$

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

- Discuss how you could improve classification performance.

More about probability

- Conditional probability of three variables

- Chain rule

- Prove!

Independence vs zero correlation

- Independence vs Pearson correlation coefficient $\rho = 0$
  
  If X and Y are independent, $\rho_{XY} = 0$.
  
  The converse is not true.

Optimisation problems we’ve seen

- Bayes decision rule (MAP decision rule)

- K-NN classification

- K-means clustering

- Dimensionality reduction to 2D with PCA

Optimisation problems: other examples

- Find the shortest path between Edinburgh and London

- Find the cheapest flights from Edinburgh to Tokyo

- For UG4 projects, find the optimal allocation of supervisors and students under given constraints (e.g. no supervisors can take more than five students.)

Types of optimisation problems

- Continuous vs Discrete optimisation

- Unconstrained vs Constrained optimisation
Continuous & unconstrained optimisation problems

Minimisation of **objective function**

\[
\min_x f(x) \quad \text{where } x \in \mathbb{R}^D, \ f : \mathbb{R}^D \to \mathbb{R}
\]

Optimal solution, \( x^* : f(x^*) \leq f(x) \) for all \( x \in \mathbb{R}^D \), satisfies \( \frac{\partial f(x)}{\partial x_i} = 0, \) for \( i = 1, \ldots, D \).

Vector representation:

\[
\nabla f(x) = \left( \frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_D} \right)^T = 0
\]

where \( 0 = (0, \ldots, 0)^T \)

---

Optimisation of a quadratic function of one variable

Optimisation problem:

\[
\min_x f(x) = ax^2 + bx + c, \quad a > 0
\]

- Approach 1:
  \[
  ax^2 + bx + c = a \left( x - \frac{b}{2a} \right)^2 - \frac{b^2}{4a} + c
  \]

- Approach 2:
  \[
  \frac{\partial f(x)}{\partial x} = 2ax + b = 0
  \]

Solution:

\[
x = -\frac{b}{2a}
\]

**Least square error line fitting**

Optimisation problem:

\[
\min_{\hat{y}_n} \frac{1}{N} \sum_{n=1}^{N} (\hat{y}_n - y_n)^2
\]

\[
\hat{y}_n = ax_n + b
\]

\[
\frac{\partial E}{\partial a} = \frac{2}{N} \sum_{n=1}^{N} (ax_n + b - y_n)x_n = 0
\]

\[
\frac{\partial E}{\partial b} = \frac{2}{N} \sum_{n=1}^{N} (ax_n + b - y_n) = 0
\]

\( \Rightarrow \) See the lecture note for details.

**Iterative optimisation**

Many optimisation problems do not have a closed-form solution! (e.g., K-means clustering)

**Iterative optimisation method**

1. Step 1: Choose an initial point \( x_0 \), and make \( t = 0 \).
2. Step 2: Choose \( x_{t+1} \) based on an update formula for \( x_t \).
3. Step 3: \( t \leftarrow t + 1 \) and go to step 2 unless stopping criterion is met.

Example of iterative optimisation methods

- **Gradient descent**
  \[
x_{t+1} = x_t - \eta \nabla f(x)|_{x=x_t}
\]
  where \( \eta > 0 \)

- Conjugate gradient method
- Newton’s method

**Gradient descent**

\[
x_{t+1} = x_t - \eta \nabla f(x)|_{x=x_t}
\]

where \( \eta > 0 \)

**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
  - Pattern recognition as optimisation problem
  - Most of optimisation problem does not have a closed-form solution \( \Rightarrow \) Iterative optimisation method
  - Check the examples in slides, and try the exercises in Note 5.