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)

Warming up

- Throwing two dice
  - Probability of (1,1)?
  - Probability of (2,5)?
- Drawing two cards from a deck of cards
  - Probability of (Club, Spade)?
  - Probability of (Club, Club)?

Warming up (cont.)

- 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) \( \frac{1}{6} \)?

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_n} )</td>
</tr>
<tr>
<td>( Y )</td>
<td>( {y_1, y_2, \ldots, y_m} )</td>
</tr>
</tbody>
</table>

Product Rule:

\[ P(Y = y_j | X = x_i) = P(Y = y_j | X = x_i) \cdot P(X = x_i) \]
\[ = P(X = x_i | Y = y_j) \cdot P(Y = y_j) \]

Abbreviation:

\[ P(Y, X) = P(Y | X) \cdot P(X) \]
\[ = P(X | Y) \cdot P(Y) \]

\( X \) and \( Y \) are independent if:

\[ P(X, Y) = P(X) \cdot P(Y) \]
\[ P(X|Y) = P(X), \quad P(Y|X) = P(Y) \]

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:

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

Example: determining the sex of fish

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

Relative frequencies of fish length

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

Possible decision boundary

<table>
<thead>
<tr>
<th>Length / cm</th>
<th>Relative frequency of male fish</th>
<th>Relative frequency of female fish</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>5</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>10</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>15</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>20</td>
<td>0.0</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**Fish questions**

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

**Fish questions (cont.)**

- How to obtain $P(Y | x)$? (where $Y = \{F, M\}$)
  - The product rule:
    \[
P(Y, x) = P(Y | x) P(x) = P(x | Y) P(Y)
    \]
  - Posterior probabilities:
    \[
P(Y | x) = \frac{P(x | Y) P(Y)}{P(x)}
    \]
    i.e.
    \[
P(M | x) = \frac{P(x | M) P(M)}{P(x)} \propto P(x | M) P(M)
    \]
    \[
P(F | x) = \frac{P(x | F) P(F)}{P(x)} \propto P(x | F) P(F)
    \]

**Bayes’ Theorem**

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

Thomas Bayes (?) (1702? – 1761)

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

c.f. Bayesian inference, Bayesian

**Bayes’ decision rule**

Class $C = \{1, \ldots, K\}$, $C_0$ to denote $C = k$; input features $X = x$

Choose the most probable class: (maximum posterior class)

\[
\hat{k}_{\text{posterior}} = \arg \max_{k \in C} P(C_k | x) = \arg \max_{k \in C} P(x | C_k) P(C_k)
\]

\[
P(C_k | x) = \frac{P(x | C_k) P(C_k)}{P(x)} = \frac{P(x | C_k) P(C_k)}{\sum_{j=1}^{K} P(x | C_j) P(C_j)}
\]

- It is known this decision rule gives minimum error rate.
  (We will discuss this in Lecture 10)

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

**Inferring labels for $x = 11$**

- Equal prior probabilities:
  \[
P(M | x = 11) = \frac{P(x = 11 | M) P(M)}{P(x = 11)}
  = \frac{P(x = 11 | M) P(M) + P(x = 11 | F) P(F)}{1.04 - 0.5}
  = 0.583
  \]
  \[
P(F | x = 11) = \frac{P(x = 11 | F) P(F)}{P(x = 11)}
  = \frac{0.14 - 0.5 + 0.10 - 0.5}{0.14 - 0.5 + 0.10 - 0.5}
  = 0.416
  \]
  \[\Rightarrow \text{classify it as male}\]

NB: For classification, no need to calculate $P(x = 11)$.
What is the value of $P(M|x=11)$?

- Equal prior probabilities:
  \[
P(M|x=11) = P(x=11|M) P(M) = 0.364 = 1.4
  \]
- Twice as many females as males: (i.e., $P(M=1)=1/3, P(F=2)=2/3$)
  \[
P(M|x=11) = P(x=11|M) P(M) = 0.14 / 1/3 = 0.7
  \]

Classify it as female.

More questions

- Assume $P(M)=P(F)=0.5$
  - What is the value of $P(M|X=4)$?
  - What is the value of $P(F|X=8)$?
  - You observe data point $x=22$.
    - To which class should it be assigned?

Discuss how you could improve classification performance.

- 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 $P(C|x)$ directly from data, right?

Likelihood vs posterior probability

\[
P(C_k|x) = \frac{P(x|C_k)P(C_k)}{\sum_{j=1}^k P(x|C_j)P(C_j)}
\]

Likelihoods vs posterior probabilities (cont.)

\[
P(C_k|x) = \frac{P(x|C_k)P(C_k)}{\sum_{j=1}^k P(x|C_j)P(C_j)}
\]

Optimisation problems

- Bayes decision rule (MAP decision rule)
  \[
k_{\text{max}} = \arg \max_{k \in C} P(C_k|x)
\]
- K-NN classification
  \[
c(z) = \arg \max_{j \in (1...C)} \sum_{i \in U_z(x)} \delta_{x_i}
\]
  where $U_z(x)$ is the set of $K$ nearest training examples to $x$.
- K-means clustering
  \[
  \min_{m} E \| m_i \|^2
  \]
  where $E = \frac{1}{N} \sum_{i=1}^{N} \sum_{x_i \in C_k} \| x_i - m_i \|^2$
- Continuous vs Discrete optimisation
- Unconstrained vs Constrained optimisation
Continuous & unconstrained optimisation problems

Minimisation of objective function

$$\min f(x) \quad \text{where } x \in \mathbb{R}^D, f: \mathbb{R}^D \to \mathbb{R}$$

Optimal solution, $$x^* : f(x^*) \leq f(x)$$ for all $$x \in \mathbb{R}^D$$, satisfies 

$$\frac{\partial f(x)}{\partial x_i} = 0, \text{ for } i = 1, \ldots, D$$

Vector representation:

$$\nabla f(x) = \left( \frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_D} \right) = 0$$

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

Optimisation of a quadratic function of one variable

Optimisation problem:

$$\min f(x)$$

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

• 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: $$x = -\frac{b}{2a}$$

Optimisation of a quadratic function of two variables

Optimisation problem:

$$\min g(x, y)$$

$$g(x, y) = ax^2 + by^2 + cxy + dx + ey + f$$

where $$a > 0, b > 0, c^2 < 4ab$$

$$\Rightarrow \quad \frac{\partial g}{\partial x} = 2ax + cy + d = 0$$

$$\frac{\partial g}{\partial y} = 2by + cx + e = 0$$

$$\left( \begin{array}{c} 2a \\ c \\ 2b \end{array} \right) \left( \begin{array}{c} x \\ y \end{array} \right) = \left( \begin{array}{c} -d \\ -e \end{array} \right)$$

Least square error line fitting

Optimisation problem:

$$\min \frac{1}{N} \sum_{n=1}^{N} (\hat{y}_n - y_n)^2$$

$$\hat{y}_n = ax_n + b$$

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

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

⇒ See the lecture note for details.

Exercise:

Optimisation problem

$$\min \frac{1}{N} \sum_{n=1}^{N} (\hat{y}_n - x_n)^2$$

$$\hat{y}_n = c x_n + d$$

Find the solution

Iterative optimisation

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

Iterative optimisation method

Step 1: Choose an initial point $$x_0$$, and make $$t = 0$$.

Step 2: Choose $$x_{t+1}$$ based on an update formula for $$x_t$$.

Step 3: $$t \leftarrow t + 1$$ and go to step 2 unless stopping criterion is met.

Example of iterative optimisation methods

• Gradient descent

$$x_{t+1} = x_t - \eta \nabla f(x)|_{x=x_t} \quad \text{where } \eta > 0$$

• Conjugate gradient method

• Newton’s method

Gradient descent

$$x_{t+1} = x_t - \eta \nabla f(x)|_{x=x_t} \quad \text{where } \eta > 0$$

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 ⇒ Iterative optimisation method

• Check the examples in slides, and try the exercises in Note 5.

Things to consider:

- Choice of $$\eta$$ (i.e. learning parameter)
- Local-minimum problem