Probability, Entropy, and Inference

Based on *David J.C. MacKay:*
Information Theory, Inference and Learning Algorithms, 2003
Chapter 2

Juha Raitio
juha.raitio@iki.fi

5th February 2004

Outline

1. On notation of probabilities
2. Meaning of probability
3. Forward and inverse probabilities
4. Probabilistic inference
5. Shannon information and entropy
6. On convexity of functions
7. Exercises

Ensembles and probabilities

- *Ensemble* X is a triple $(x, \mathcal{A}_X, \mathcal{P}_X)$, where
  - x is the *outcome* of random variable
  - $\mathcal{A}_X = \{a_1, a_2, \ldots, a_I\}$ are the *possible values* for x
  - $\mathcal{P}_X = \{p_1, p_2, \ldots, p_I\}$ are the *probabilities* of outcomes $P(x = a_i) = p_i$
  - $p_i \geq 0$
  - $\sum_{a_i \in \mathcal{A}_X} P(x = a_i) = 1$
- $P(x = a_i)$ may be written as $P(a_i)$ or $P(x)$
- Probability of a subset T of $\mathcal{A}_x$

$$P(T) = P(x \in T) = \sum_{a_i \in T} P(x = a_i) \quad (1)$$

Joint ensembles and marginal probabilities

- *Joint ensemble* XY
  - *Outcome* is an ordered pair x, y (or xy)
  - *Possible values* $\mathcal{A}_X = \{a_1, a_2, \ldots, a_I\}$ and $\mathcal{A}_Y = \{b_1, b_2, \ldots, b_J\}$
  - *Joint probability* $P(x, y)$
- *Marginal probabilities*

$$P(x = a_i) \equiv \sum_{y \in \mathcal{A}_Y} P(x = a_i, y) \quad (2)$$

$$P(y) \equiv \sum_{x \in \mathcal{A}_X} P(y, x) \quad (3)$$

Conditioning rules

- Conditional probability

$$P(x = a_i|y = b_j) \equiv \frac{P(x = a_i, y = b_j)}{P(y = b_j)}, \quad P(y = b_j) \neq 0 \tag{4}$$

- Assumptions $\mathcal{H}$
  - ”the probability that x equals a_i, given $\mathcal{H}$”
- *Product (chain) rule*

$$P(x, y|\mathcal{H}) = P(x|y, \mathcal{H})P(y|\mathcal{H}) = P(y|x, \mathcal{H})P(x|\mathcal{H}) \tag{5}$$

- *Sum rule*

$$P(x|\mathcal{H}) = \sum_y P(x, y|\mathcal{H}) = \sum_y P(x|y, \mathcal{H})P(y|\mathcal{H}) \tag{6}$$

Bayes theorem and independence

- *Bayes theorem*

$$\begin{aligned} P(y|x, \mathcal{H}) &= \frac{P(x|y, \mathcal{H})P(y|\mathcal{H})}{P(x|\mathcal{H})} && (7) \\ &= \frac{P(x|y, \mathcal{H})P(y|\mathcal{H})}{\sum_{y'} P(x|y', \mathcal{H})P(y'|\mathcal{H})} && (8) \end{aligned}$$

- Two random variables X and Y are *independent* ($X \perp Y$) if and only if

$$P(x, y) = P(x)P(y) \tag{9}$$

Two meanings for probability

- *Frequentist* view of probability
  - Probabilities are *frequencies of outcomes in random experiments*
  - Probabilities describe random variables
- *Bayesian* view of probability
  - Probabilities are *degrees of belief* in propositions
  - Probabilities describe assumptions, and inferences given assumptions
  - *Subjective intepretation of probability*
    “you cannot do inference without making assumptions”

Forward and inverse probabilities

- Assume *generative model* describing a process giving rise to some data
- *Forward probability*
  - Task is to compute probability distribution of some quantity that depends on data
- *Inverse probability*
  - Task is to compute probability distribution of *unobserved variables* given data
  - Requires use of Bayes’ theorem

Inference with inverse probabilities

- Inference on parameters θ given data D and hypothesis $\mathcal{H}$ by Bayes' theorem

$$P(\theta|D,\mathcal{H}) = \frac{P(D|\theta,\mathcal{H})P(\theta|\mathcal{H})}{P(D|\mathcal{H})}, \tag{10}$$

where

$P(\theta|\mathcal{H})$ is the *prior probability* for parameters
$P(D|\theta,\mathcal{H})$ is the *likelihood* of the parameters given the data
$P(D|\mathcal{H})$ is the *evidence*
$P(\theta|D,\mathcal{H})$ is the *posterior probability* for parameters

- in written

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \tag{11}$$

Shannon information and entropy

- *Shannon information content of an outcome* $x = a_i$ (bits)

$$h(x = a_i) = \log_2 \frac{1}{P(x = a_i)} \tag{12}$$

- *Entropy of an ensemble* X (bits)

$$H(X) \equiv \sum_{x \in \mathcal{A}_X} P(x) log \frac{1}{P(x)} \tag{13}$$

- *Joint entropy* of X, Y

$$H(X,Y) \equiv \sum_{xy \in \mathcal{A}_X \mathcal{A}_Y} P(x,y) log \frac{1}{P(x,y)} \tag{14}$$

Decomposability of entropy

- Entropy of probability distribution $\mathbf{p} = \{p_1, p_2, \ldots, p_I\}$

$$H(\mathbf{p}) = H(p_1, 1-p_1) + (1-p_1) H\left(\frac{p_2}{1-p_1}, \frac{p_3}{1-p_1}, \ldots, \frac{p_I}{1-p_1}\right) \tag{15}$$

- More generally

$$\begin{aligned} H(\mathbf{p}) \quad = \quad & H[(p_1 + p_2 + \ldots + p_m), (p_{m+1} + p_{m+2} + \ldots + p_I)] \\ & + (p_1 + \ldots + p_m) H\left(\frac{p_1}{(p_1 + \ldots + p_m)}, \ldots, \frac{p_m}{(p_1 + \ldots + p_m)}\right) \\ & + (p_{m+1} + \ldots + p_I) H\left(\frac{p_{m+1}}{(p_{m+1} + \ldots + p_I)}, \ldots, \frac{p_I}{(p_{m+1} + \ldots + p_I)}\right) \end{aligned} \tag{16}$$

Relative entropy

- *Kullback-Leibler divergence* between $P(x)$ and $Q(x)$ over alphabet $\mathcal{A}_X$

$$D_{KL}(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \tag{17}$$

- Properties of relative entropy
  - *Gibbs' inequality*: $D_{KL}(P\|Q) \geq 0 \quad \text{and} \quad D_{KL}(P\|Q) = 0, \text{ if } P = Q$
  - in general $D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$

Convex and concave functions

- $f(x)$ is *convex* over (a, b), if for all $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \tag{18}$$

- $f(x)$ is *concave* is the above above holds for f with the inequities reversed
- $f(x)$ is *strictly convex (concave)* if the equality in (18) holds only for $\lambda = 0$ and $\lambda = 1$
- *Jensen's inequality* for convex function $f(x)$ of random variable x

$$\mathcal{E}\left[f(x)\right] \geq f\left(\mathcal{E}\left[x\right]\right), \quad \text{where}\, \mathcal{E}\, \text{denotes expectation} \tag{19}$$

- If $f(\mathbf{x})$ is convex (concave) and $\nabla f(\mathbf{x}) = 0$, then f has its minimum (maximum) value at $\mathbf{x}$

Problems

1. A circular coin of diameter a is thrown onto a square grid whose squares are $b \times b$, $(a < b)$. What is the probability that the coin will lie entirely within one square? (MacKay exercise 2.31)

2. The inhabitants of an island tell the truth one third of the time. They lie with probability $2/3$. On an occasion, after one of them made a statement, you ask another 'was the statement true?' and he says 'yes'. What is the probability that the statement was indeed true? (MacKay exercise 2.37)