IR2 – Data Mining 2002–2003

Exercises I

Peter Lucas
Institute for Computing and Information Sciences
University of Nijmegen

1 Basic probability theory

Let $\mathcal{B}$ be a Boolean algebra, and $P : \mathcal{B} \to [0, 1]$ be a probability distribution.

a. Explain by means of a Venn diagram what it means if it holds that $P(a \vee b) < P(a) + P(b)$.

b. Prove that if a is a subevent of b, $a, b \in \mathcal{B}$, then $P(a) \leq P(b)$.

c. Explain in words why $P(a \mid a) = 1$, $\forall a \in \mathcal{B}$.

d. *Sampling without replacement.* Consider a room with 10 computers with Windows XP, of which 3 are unable to boot. If you try to use two successive computers at random, what is the probability that you can actually use both of them? (Hint: use Bayes' rule.)

2 Mathematical expectation

a. Consider the function g defined as $g(X) = X^2$ for random variable X. Let X be uniformly distributed on the closed interval $[-1, 1]$. What is $\mathrm{E}(g(X))$?

b. Determine the expectation for the function h, with $h(X) = aX + b$, where the random variable X is uniformly distributed on the closed interval $[0, 1]$.

c. Prove that $\mathrm{E}(X - E(X)) = 0$.

3 Bias-variance decomposition

Consider the following functions

$$
\begin{aligned}
f(x) &= a_1x + a_0 \\
g(x) &= a_2x^2 + a_1x + a_0 \\
h(x) &= a_3x^3 + a_2x^2 + a_1x + a_0
\end{aligned}
$$

and Table 1 with results of these functions after they have been fitted using least-squared approximation to data of 1000 different training sets. The function r underlies the process that generated the data.

a. What is your opinion about the amount of bias in each individual function?

x	$r(x)$	E(f)	E(g)	E(h)	V(f)	V(g)	V(h)
1	2.50	2.48	2.48	2.49	0.34	0.61	0.84
2	3.00	2.99	2.98	2.98	0.25	0.27	0.29
3	3.50	3.49	3.49	3.48	0.18	0.18	0.33
4	4.00	3.99	4.00	3.99	0.13	0.20	0.32

Table 1: Results of experiments.

b. What is your opinion about the amount of variance in each individual function?

c. By looking at these results, what is the most likely form of the function r?

4 Discrete distributions

We carry out a number of experiments n, where x results are successful (i.e. yield event e) and $n - x$ yield failure (i.e. yield event $\neg e$). If $p = P(e)$ and $q = P(\neg e)$, then the probability that x trials are successful is:

$$f(x) = \binom{n}{x} p^x q^{n-x} = \binom{n}{x} p^x (1-p)^{n-x}$$

i.e. is described by a *binomial distribution*, also called Bernoulli distribution. A binomial distribution describes *sampling with replacement*.

If the events e concern objects that have or have not particular properties (e.g. are defective or not), and if we assume that there is a maximum of N of these objects that can be observed, of which M have that particular property. Then

$$P(e) = p = \frac{M}{N}$$

Next, consider the situation that we again carry out a number of experiments n, but now we assume that each object that is observed is deleted afterwards. The probability that out of n experiments x are successful (yield event e rather than $\neg e$) is

$$h(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

This is because:

- There are $\binom{N}{n}$ ways of picking n objects from N.
- There are $\binom{M}{x}$ ways of picking x objects from M with the same property.
- There are $\binom{N-M}{n-x}$ ways of picking $n - x$ objects that do not have the property from the total of $N - M$ objects that do not have the property either.

This is the *hypergeometric distribution* that describes *sampling without replacement.*

Reconsider exercise 1(d).

a. Use the hypergeometric distribution to compute the probability that you can use both computers.

b. Use the binomial distribution to compute the probability that you can use both computers. Compare the two results.