Exercises to Statistical Learning WS 2017/18, Series 1

Daniel Mayer
Matrikel Nr.: 2519400

October 26, 2017

1.1

Simplify the following expressions (from Gnedenko)!

(a) $(A+B)(B+C)$ by commutativity: $=(B+A)(B+C)$
simple application of the distributive law yields: $(A+B)(B+C)=B+(AC)$

(b) $(A+B)(A+\overline{B}) \overset{distributivity}{=} A+\underbrace{B\overline{B}}_{=\emptyset}=A$

(c) $(A+B)(\overline{A}+B)(A+\overline{B}) = (B+A)(B+\overline{A}(A\overline{B}) = (B+A\overline{A})(A\overline{B} = B(A+\overline{B}) = BA+B\overline{B} = AB$

1.2

determine the probabilities of getting the following results when tossing a coin 7 times.

(a) ZWZZWWZ, this sequence is uniquely so it is the probability of tossing 7 times the right thing, hence $\frac{1}{2^7}$

(b) the probability of tossing 4 heads is given by $\binom{7}{4}\left(\frac{1}{2}\right)^7$ which results in $\frac{35}{128}$

(c) the probability of tossing at least four times head, is the same probability as that of tossing at least four tails, therefore it is $\frac{1}{2}$

1.3 Bayes' formula

Lets first establish some identities:
From the formula for conditional probability

$$P(A_i|B)=\frac{P(A_iB)}{P(B)} \tag{1}$$

we solve for $P(A_iB)$

$$P(BA_i)=P(A_i|B)P(B)=P(A_iB)=P(B|A_i)P(A_i) \tag{2}$$

further for

$$\bigcup A_i = \Omega$$

and

$$A_i \cap A_j = \emptyset, \quad \forall i \neq j$$

we can expand $P(B)$ to $P(\Omega|B)P(B)$, since $P(\Omega|X) = 1$ for $X \neq \emptyset$ ie:

$$P(B) = \underbrace{\sum_j P(A_j|B)}_{=1} P(B) \overset{(2)}{=} \sum_j P(B|A_j)P(A_j) \tag{3}$$

it is left to substitute the respective terms of (2) and (3) in equation (1)

$$P(A_i|B) = \frac{P(B|A_i)P(A_j)}{\sum_i P(B|A_j)P(A_j)} \tag{4}$$

1.4

We have the sensitivity $P(T^+|D^+) = 0.98$ specificity $P(T^-|D^-) = 0.98$ and the prevalence $P(D^+) = 0.01$ given.
It is further straightforward to calculate the opposites $P(D^-) = 1 - P(D^+) = 0.99$ and $P(T^+|D^-) = 1 - P(T|D^-) = 0.02$

(a) the PPV is

$$P(D^+|T^+) = \frac{P(T^+|D^+)P(D^+)}{P(T^+|D^+)P(D^+) + P(T^+|D^-)P(D^-)} \tag{5}$$

$$= \frac{0.98 \times 0.01}{0.98 \times 0.01 + 0.02 \times 0.99} \approx 0.33$$

(b) we have to take the PPV of the first test as the Prevalence of the second test, since of all persons with a positive first test 0.33% have a sick fetus, thus

$$P(D^+|(T^+)^2) = \frac{0.98 \times 0.33}{0.98 \times 0.33 + 0.2 \times 0.67} = 0.96$$

of course this only works if the tests are completely uncorrelatet.

1.5 Variance of the binomial distribution

recapitulate:
The probability function of the binomial distriibution is given by :

$$p_n(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

and the expectation value is $E(X) = \sum k \binom{n}{k} p^k (1-p)^{n-k} = np$
For the Variance we are looking for the expectation value of

$$(X - E(X))^2 = (k - np)^2 = k^2 - 2knp + n^2p^2$$

The Variance is found as:

$$
\begin{aligned}
V(X) &= \sum_k (k^2 - 2knp + n^2p^2)\binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum k^2 \binom{n}{k} p^k (1-p)^{n-k} + 2np \underbrace{\sum k \binom{n}{k} p^k (1-p)^{n-k}}_{=E(X)=np} + n^2p^2 \underbrace{\sum \binom{n}{k} p^k (1-p)^{n-k}}_{=1} \\
&= n^2p^2 - 2n^2p^2 + \underbrace{\sum k \binom{n}{k} p^k (1-p)^{n-k}}_{=E(X)=np} + \sum k(k-1)\binom{n}{k} p^k (1-p)^{n-k} \\
&= -n^2p^2 + np + n(n-1)p^2 \sum \binom{n-2}{k-2} p^{k-2} (1-p)^{n-k} \\
&\text{change of variables } k' = k-2, \quad n' = n-2 \\
&= -n^2p^2 + np + n(n-1)p^2 \underbrace{\sum \binom{n'}{k'} p^{k'} (1-p)^{n-k}}_{=1} \\
&= np - np^2 \\
&= np(1-p)
\end{aligned}
\tag{6}
$$

1.6 Variance of the Normal distribution

The probability density of the normal distribution is given by

$$\frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-a)^2}{2\sigma^2})$$

the expectation value is $E(X) = a$
The variance is:

$$V(X)\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-a)^2 \exp(-\frac{(x-a)^2}{2\sigma^2}) \tag{7}$$

we substitute $y = \frac{x-a}{\sigma}$ thus:

$$x = \sigma y + a, \quad dx = \sigma dy \tag{8}$$

$$V(X) = \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-\frac{y^2}{2}} \sigma dy \tag{9}$$

This can be solved by partial integration ie by using the fact that:

$$uv' = (uv)' - u'v$$

we identify $ye^{-\frac{y^2}{2}}$ as being more easily integrable, so we write:

$$\int_{-\infty}^{\infty} y \left(ye^{-\frac{y^2}{2}}\right) dy = \underbrace{\left[y(-e^{-\frac{y^2}{2}})\right]_{-\infty}^{\infty}}_{\text{odd thus }=0} - \underbrace{\int_{-\infty}^{\infty} -e^{-\frac{y^2}{2}} dy}_{=\sqrt{2\pi}}$$

Thus the variance of the normal function is simply σ^2 which isn't just a fancy coincidence.