Departamento de Ingeniería Informática

Practical Lesson

Introduction to Probabilities for Machine Learning

Data Science I

Cristóbal Donoso

5 Septiembre 2018

Problem 1

My neighbor has two children. Assuming that the gender of a child is like a coin flip, it is most likely, a priori, that my neighbor has one boy and one girl, with probability $1/2$. The other possibilities -two boys or two girls- have probabilities $1/4$ and $1/4$.

a. Suppose I ask him whether he has any boys, and he says yes. What is the probability that one child is a girl?
b. Suppose instead that I happen to see one of his children run by, and it is a boy. What is the probability that the other child is a girl?

Problem 2

Show that the variance of a sum is $\mathrm{VAR}[X+Y] = \mathrm{VAR}[X] + \mathrm{VAR}[Y] + 2\mathrm{cov}[X,Y]$ where $\mathrm{cov}[X,Y]$ is the covariance between X and Y

Problem 3

Let H $\in \{1, ..., K\}$ be a discrete random variable, and let e_1 and e_2 be the observed values of two other random variables E_1 and E_2. Suppose we wish to calculate the vector

$$\overrightarrow{P}(e_1, e_2) = (P(H = 1|e_1, e_2), ..., P(H = K|e_1, e_2)) \tag{0.15}$$

Which of the following sets of numbers are sufficient for the calculation ?

a. i. $P(e_1, e_2), P(H), P(e_1|H), P(e_2|H)$
   ii. $P(e_1, e_2), P(H), P(e_1, e_2|H)$
   iii. $P(e_1|H), P(e_2|H), p(H)$
b. Now suppose we now assume $E_1 \perp E_2|H$ (i.e E_1 and E_2 are conditionally independent given H), Which of the above 3 sets are sufficient now ?

Show your calculations as well as giving the final results. Hint : use Bayes rule

Problem 4

After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease and that the test is 99% accurate (i.e., the probability of testing positive when you do have the disease is 0.99, as is the probability of testing negative when you don't have the disease). The good news is that it is a rare disease, striking only 1 in 10,000 people of your age. What is the probability that you actually have the disease?

Problem 5

Verify that the Bernoulli distribution

$$Bern(x|\mu) = \mu^x(1-\mu)^{1-x} \tag{0.27}$$

satisfies the following properties

$$\sum_{x=0}^{1} p(x|\mu) = 1 \tag{0.28}$$

$$E[x] = \mu \tag{0.29}$$

$$var[x] = \mu(1-\mu) \tag{0.30}$$

Problem 6

Consider the generalization of the squared loss function for a single target variable t

$$E[L] = \int\int \{y(x) - t\}^2 p(x,t) dx dt \tag{0.36}$$

to the case of multiple target variables described by the vector $\mathbf{t}$ given by

$$E[L(\overrightarrow{t}, y(x))] = \int\int ||y(x) - \overrightarrow{t}||^2 p(x, \overrightarrow{t}) dx d\overrightarrow{t} \tag{0.37}$$

Using the calculus of variations, show that the function $y(x)$ for which this expected loss is minimized is given by $y(x) = E_t[\overrightarrow{t}|x]$. Show that this result reduces to

$$y(x) = \frac{\int tp(x,t)dt}{p(x)} = \int tp(t|x)dt = E_t[t|x] \tag{0.38}$$

for the case of a single target variable t