Answers to Machine Learning Exercises

Björn Lindqvist
bjourne@gmail.com

February 27, 2018

About

LaTeXsolutions to exercises in machine learning.

Contents

Introduction to Statistical Learning

Chapter 2.4

Exercise 1

a) A flexible method is better because the large sample size protects against overfitting.

b) Creating a model for this scenario is very hard because of the large number of predictors. Again I believe an inflexible method would be most appropriate because of overfitting risks.

c) A flexible method must be used to accurately predict non-linear relationships.

d) An inflexible method is preferrable. A flexible one would be *fooled by randomness* and increase variance.

Exercise 2

a) Salary is best modelled as a non-discreete variable and we are therefore dealing with a regression problem. We want to understand the relationship between variables and it is therefore an inference problem.

b) It is a classification problem because outcomes are classified into *success* or *failure*. It is mostly about prediction because we want to estimate the likelihood of a future event. It might also be interesting to understand what variables affect whether a product succeeds or not. Therefore it can also be seen as an inference problem.

c) This is a pure prediction problem as we are unlikely to be able to understand what factors affect the % change in the USD/Euro exchange rate. It is also a regression problem as % change is a continuous variable.

Exercise 7

a) Distances are 3.0, 2.0, 3.16, 2.24, 1.41 and 1.74.

b) With $K = 1$ class is *green*.

c) With $K = 3$ class is drawn from points $\{2, 5, 6\}$ with classes $\{red, green, red\}$ and is therefore *red*.

d) If the Bayes decision boundary is highly non-linear, then the *best* value for K would be small. The larger the value for K, the smoother the decision boundary.

Computer vision: models, learning and inference

Chapter 2

2.1 Give a real-world example of a joint distribution $Pr(x, y)$ where x is discrete and y is continuous.

The x variable can represent height in centimeters and y European shoe size.

2.2 What remains if I marginalize a joint distribution $Pr(v, w, x, y, z)$ over five variables with respect to variables w and y? What remains if I marginalize the resulting distribution with respect to v?

Marginalization over the variables w and y results in the joint distribution $Pr(v, x, z)$. Marginalizing the resulting distribution over v results in $Pr(x, z)$.

2.3 Show that the following relation is true:

$$Pr(w, x, y, z) = Pr(x, y)Pr(z|w, x, y)Pr(w|x, y)$$

Expansion of the L.H.S:

$$\begin{aligned} Pr(w, x, y, z) &= Pr(z|w, x, y)Pr(w, x, y) \\ &= Pr(z|w, x, y)Pr(w|x, y)Pr(x, y) \end{aligned}$$

2.4 In my pocket there are two coins. Coin 1 is unbiased, so the likelihood $Pr(h = 1|c = 1)$ of getting heads is 0.5 and the likelihood $Pr(h = 0|c = 1)$ of getting tails is also 0.5. Coin 2 is biased, so the likelihood $Pr(h = 1|c = 2)$ of getting heads is 0.8 and the likelihood $Pr(h = 0|c = 2)$ of getting tails is 0.2. I reach into my pocket and draw one of the coins at random. There is an equal prior probability I might have picked either coin. I flip the coin and observe a head. Use Bayes' rule to compute the posterior probability that I chose coin 2.

The posterior probability is expressed as $Pr(y|x)$, the probability of y given the evidence x. In this case, we are calculating $Pr(c = 2|h = 1)$, the probability that coin 2 was choosen given that we observed a heads. We apply Bayes' rule with y set to $c = 2$ and x set to $h = 1$

$$\begin{aligned} Pr(c = 2|h = 1) &= \frac{Pr(h = 1|c = 2)Pr(c = 2)}{Pr(h = 1)} \\ &= \frac{0.8 \cdot 0.5}{Pr(h = 1)} \end{aligned}$$

The exercise has already given us the probabilities for $Pr(h = 1|c = 2)$ and $Pr(c = 2)$, so what is left is calculating $Pr(h = 1)$ by marginalizing with

respect to c:

$$\begin{aligned} Pr(h=1) &= \sum_{i=1}^{2} Pr(h=1, c=i) \\ &= Pr(h=1, c=1) + Pr(h=1, c=2) \\ &= Pr(h=1|c=1)Pr(c=1) + Pr(h=1|c=2)Pr(c=2) \\ &= 0.5 \cdot 0.5 + 0.8 \cdot 0.5 \\ &= 0.65 \end{aligned}$$

The value inserted in the original equation results in the answer about 62 %.

2.5 If variables x and y are independent and variables x and z are independent, does it follow that variables y and z are independent?

No. Let x and y be the result of fair coin tosses, with heads represented as 1 and tails as 0. Then x and y are independent variables. Let z be defined as $z = 3y$. Then x and z are independent, but y and z are very much dependent.

2.6 Use the following equation

$$Pr(x|y=y^*) = \frac{Pr(x, y=y^*)}{\int Pr(x, y=y^*)dx} = \frac{Pr(x, y=y^*)}{Pr(y=y^*)}$$

to show that when x and y are independent, the marginal distribution $Pr(x)$ is the same as the conditional distribution $Pr(x|y=y^*)$ for any y^*.

For independent variables we have $Pr(x, y) = Pr(x)Pr(y)$ therefore:

$$Pr(x|y=y^*) = \frac{Pr(x, y=y^*)}{Pr(y=y^*)} = \frac{Pr(x)Pr(y=y^*)}{Pr(y=y^*)} = Pr(x)$$

2.7 The joint probability $Pr(w, x, y, z)$ over four variables factorizes as

$$Pr(w, x, y, z) = Pr(w)Pr(z|y)Pr(y|x, w)Pr(x)$$

Demonstrate that x is independent of w by showing that $Pr(x, w) = Pr(x)Pr(w)$.

Ask for help.

2.7 Consider a biased die where the probabilities of rolling sides 1, 2, 3, 4, 5, 6 are 1/12, 1/12, 1/12, 1/12, 1/6, 1/2, respectively. What is the expected value of the die? If I roll the die twice, what is the expected value of the sum of the two rolls?

The expected value of a function of a discreete random variable is defined as

$$E(f(x)) = \sum_x f(x)Pr(x)$$

, so we simply plug in the numbers

$$\begin{aligned} E(f(x)) &= 1 \cdot \tfrac{1}{12} + 2 \cdot \tfrac{1}{12} + 3 \cdot \tfrac{1}{12} + 4 \cdot \tfrac{1}{12} + 5 \cdot \tfrac{1}{6} + 6 \cdot \tfrac{1}{2} \\ &= 10 \cdot \tfrac{1}{12} + 5 \cdot \tfrac{1}{6} + 6 \cdot \tfrac{1}{2} = \tfrac{20}{12} + \tfrac{36}{12} = \tfrac{56}{12} = \tfrac{14}{3} \end{aligned}$$

2.9 Prove the four relations for manipulating expectations.

$$\begin{aligned} E(k) &= k \\ E(kf(x)) &= kE(f(x)) \\ E(f(x)+g(x)) &= E(f(x)) + E(g(x)) \\ E(f(x)g(y)) &= E(f(x))E(g(y)) \end{aligned}$$

for the last case, it is assumed that x and y are independent so you will need to use the definition of independence.

We prove the first relation using a continuous random variable

$$E(k) = \int kPr(x)\,dx = k\int Pr(x)\,dx = k$$

since we know that integrating over a continuous variable always yields one. For the second relation

$$E(kf(x)) = \int kf(x)Pr(x)\,dx = k\int f(x)Pr(x)\,dx = kE(f(x))$$

The third relation

$$\begin{aligned} E(f(x)+g(x)) &= \int (f(x)+g(x))Pr(x)\,dx \\ &= \int f(x)Pr(x)\,dx + \int g(x)Pr(x)\,dx \\ &= E(f(x)) + E(g(x)) \end{aligned}$$

The definition of independence is

$$Pr(x,y) = Pr(x|y)Pr(y) = Pr(x)Pr(y)$$

giving us

$$\begin{aligned} E(f(x)g(y)) &= \iint f(x)g(y)Pr(x,y)\,dx\,dy \\ &= \iint f(x)g(y)Pr(x)Pr(y)\,dx\,dy \end{aligned}$$

The integrand is fortunately separable

$$\iint f(x)g(y)Pr(x)Pr(y)\,dx\,dy = \int f(x)Pr(x)\,dx \cdot \int g(y)Pr(y)\,dy$$

Which is the same as $E(f(x))E(g(y))$.

Chapter 3

3.1 Consider a variable x which is Bernoulli distributed with parameter λ. Show that the mean $E(x)$ is λ and the variance $E((x-E(x))^2)$ is $\lambda(1-\lambda)$.

We begin by showing the mean

$$E(x) = \sum_{x=0}^{1} f(x)P(x) = 0 \cdot (1-\lambda) + 1 \cdot \lambda = \lambda$$

and then the variance using the formula $Var(x) = E(x^2) - E(x)^2$

$$E((x - E(x))^2) = E(x^2) - E(x)^2$$

first term

$$E(x^2) = \sum_{x=0}^{1} f(x^2)P(x^2) = 0 \cdot (1-\lambda) + 1 \cdot \lambda = \lambda$$

second term

$$E(x)^2 = E(x)E(x) = \lambda^2$$

giving us the result $\lambda - \lambda^2$. A smarter method, using the identities $E(x) = E(x^2) = \lambda$ and $E(k) = k$

$$\begin{aligned} E((x - E(x))^2) &= E(x^2 + E(x)^2 - 2xE(x)) \\ &= E(x^2) + E(x)^2 - E(2xE(x)) \\ &= \lambda + \lambda^2 - E(2x\lambda) \\ &= \lambda + \lambda^2 - 2\lambda\lambda \\ &= \lambda - \lambda^2 = \lambda(1-\lambda) \end{aligned}$$

Exams

Exam 2017-10-21

Explain the following concepts:

maximum a posteriori estimation What is it