Statistical Machine Learning 2018

Exercises and answers, week 4

28 September 2018

TUTORIAL

Exercise 1

A factory produces products X. 75% is of quality $x = 1$ and the remainder of quality $x = 2$. There is a test Z, which can be a real number z between 0 and 1. The conditional probability density of z, depending on the quality x is

$$\begin{aligned} p(z|x=1) &= 2(1-z) \\ p(z|x=2) &= 1 \end{aligned}$$

1. Interpret these equations and compute $p(x|z)$ using Bayes' rule

   ANSWER:
   $$p(x=1|z) = \frac{0.75(2-2z)}{0.75(2-2z)+0.25} = \frac{6-6z}{7-6z}$$
   and
   $$p(x=2|z) = \frac{1}{7-6z}$$

2. Compute the Bayes optimal decision to minimize misclassification rate as function of z, i.e. for which z should one classify $x = 1$ and for which z should one classify $x = 2$.

   ANSWER: Decision boundary is where $p(x = 1|z) = p(x = 2|z)$, so if $6 - 6z = 1$ i.e., $z = 5/6$. Classify $x = 1$ if z smaller than $5/6$ (since then $p(x = 1|z) > p(x = 2|z)$ and $x = 2$ if z is larger. If $z = 5/6$, it does not matter.

3. Suppose we have a loss matrix L_{kj}, expressing the loss for classifying as $x = j$ while the true class is k. Suppose this matrix is given by

   $$L_{11} = L_{22} = 0, \quad L_{12} = 1, \quad L_{21} = 5$$

   Compute the optimal decision boundary to minimize expected loss.

   ANSWER: When classifying a point with test value z as $x = j$, the expected loss (given z) is given by $\mathbb{E}[L_{\cdot j}|z] = \sum_k p(x = k|z)L_{kj}$. In words, this conditional expected loss is the *weighted average* of the losses for classifying as $x = j$ while the true class is $k = 1, 2$, weighted with the posterior probability of the true class being k, given that the test result is z. The decision boundary is now given by those values of z for which the expected loss when classifying as $x = 1$ equals the expected loss when classifying as $x = 2$, i.e., for which $\mathbb{E}[L_{\cdot 1}|z] = \mathrm{E}[L_{\cdot 2}|z]$. The decision boundary is therefore defined by the equation

   $$p(x=1|z)L_{11} + p(x=2|z)L_{21} \quad = \quad p(x=1|z)L_{12} + p(x=2|z)L_{22}.$$

Since $L_{11} = L_{22} = 0$, this simplifies to $L_{12}p(x = 1|z) = L_{21}p(x = 2|z)$. Substituting the posterior probabilities calculated earlier, we arrive at $6 - 6z = 5$, so $z = 1/6$. Classify $x = 1$ if z smaller than $1/6$ (since then $L_{12}p(x = 1|z) > L_{21}p(x = 2|z)$, i.e. the loss of classifying as 2 is larger) and classify as $x = 2$ if z is larger than $1/6$.

Exercise 2

A factory produces products X. 80% is of quality $x = 1$ (functioning) and the remainder of quality $x = 2$ (faulty). Before shipping the product to the customer, an employee tests if it is functioning or faulty. We call this test Y and note that it can return either 1 (functioning) or 2 (faulty).

Products marked functioning by the test will be shipped immediately. Products marked faulty will first be sent to the repair center in order to be fixed and then will be shipped. Since faulty products need to be repaired anyway, we do not consider it a loss to send them to the repair center. However, the repair cost of 1 euro is wasted on a functioning product. Shipping faulty products, on the other hand, incurs a (potentially) large loss of c euros due to damages claims.

1. From the information in the text, derive the loss matrix $\mathbf{L}$, which consists of elements L_{kj} expressing the loss for classifying as $x = j$ while the true class is k.

   ANSWER: As discussed in the text, there is no loss for correct classification, so $L_{11} = L_{22} = 0$. On the other hand, the loss for classifying a functioning product as faulty is given by the repair cost of 1 euro, which means that $L_{12} = 1$. The loss for classifying a faulty product as functioning is given by the damages claim of c euros, so $L_{21} = c$. In conclusion, the loss matrix is:

$$\mathbf{L} = \begin{pmatrix} 0 & 1 \\ c & 0 \end{pmatrix}.$$

2. The test Y correctly classifies 60% of the functioning products and 90% of the faulty ones. Derive the joint probability distribution $p(x, y)$ from all the information given, then compute the expected loss using the loss matrix found above.

   ANSWER:

   Given the information in the text, we know the probabilities $p(y|x)$ and $p(x)$:

$$\begin{aligned} p(x=1) &= 0.8 & p(x=2) &= 0.2 \\ p(y=1|x=1) &= 0.6 & p(y=2|x=1) &= 0.4 \\ p(y=1|x=2) &= 0.1 & p(y=2|x=2) &= 0.9 \end{aligned}$$

   Computing the joint probability is then straightforward:

$$\begin{aligned} p(x=1,y=1) &= p(x=1)p(y=1|x=1) = 0.48 & p(x=1,y=2) &= p(x=1)p(y=2|x=1) = 0.32 \\ p(x=2,y=1) &= p(x=2)p(y=1|x=2) = 0.02 & p(x=2,y=2) &= p(x=2)p(y=2|x=2) = 0.18 \end{aligned}$$

   The expected loss of making a decision based on test Y is given by:

$$\begin{aligned} \mathbb{E}[L] &= L_{11}p(x=1,y=1) + L_{12}p(x=1,y=2) + L_{21}p(x=2,y=1) + L_{22}p(x=2,y=2) \\ &= 0 \cdot 0.48 + 1 \cdot 0.32 + c \cdot 0.02 + 0 \cdot 0.18 \\ &= \frac{16+c}{50} \end{aligned}$$

3. Derive the conditional probability distribution $p(x|y)$. With test Y, which will be the more commonly incurred loss: unnecessary repairs or damages claims?

   ANSWER: To obtain the probability of a product being functioning or faulty given the result of test Y, we could either use Bayes' rule, i.e. $p(x|y) = \dfrac{p(y|x)p(x)}{\sum_x p(y|x)p(x)}$, or, since we have already computed the joint probability distribution, we can directly employ the sum and product rules of probability, i.e. $p(x|y) = \dfrac{p(x,y)}{\sum_x p(x,y)}$. Note that the two approaches are equivalent. Numerically, we get:

   $$p(x=1|y=1) = \frac{24}{25} \qquad p(x=2|y=1) = \frac{1}{25}$$
   $$p(x=1|y=2) = \frac{16}{25} \qquad p(x=2|y=2) = \frac{9}{25}$$

   Based on these posterior probabilities, we can see that it is more likely to incur a loss due to unnecessary repairs ($p = \frac{16}{25}$) than due to damages claims ($p = \frac{1}{25}$). This indicates that the test Y is suitable if we are trying to avoid the potentially large cost c of damages claims.

4. The previous test Y is an overly strict human evaluation of the product state, leading to many functioning products being sent for repairs. We wish to introduce an automatic test Z to help with the pre-shipping selection process. This new test returns a continuous value z between 0 and 1, where higher values indicate the product is more likely to be functioning. The joint probability of the state of the product and the value of z is given by:

   $$p(x=1,z) = z, \quad p(x=2,z) = 1-z.$$

   Compute the optimal decision boundary that minimizes the expected loss, given the same loss matrix above. What is the expected loss if we always make the optimal decision?

   ANSWER:

   Given a particular value of z, we can either decide that the product is functioning (assign to class 1) or faulty (assign to class 2). In the first case, the incurred loss is:

   $$L_{11}p(x=1,z) + L_{21}p(x=2,z) = cp(x=2,z) = c(1-z)$$

   In the second case, the incurred loss is:

   $$L_{12}p(x=2,z) + L_{22}p(x=1,z) = p(x=1,z) = z$$

   The optimal decision given z is the one that incurs the smaller loss. To find the decision boundary, we find the value of z where these losses are equal:

   $$c(1-z) = z \iff z = \frac{c}{c+1}.$$

   If $z \geq \dfrac{c}{c+1}$, we assign the product to class 1 (functioning), otherwise we assign it to class

2 (faulty). Given the optimal decision boundary, we can now compute the expected loss:

$$
\begin{aligned}
\mathbb{E}[L] &= \int_{\frac{c}{c+1}}^{1} [L_{11}p(x=1,z) + L_{21}p(x=2,z)]dz + \int_{0}^{\frac{c}{c+1}} [L_{12}p(x=1,z) + L_{22}p(x=2,z)]dz \\
&= c\int_{\frac{c}{c+1}}^{1} p(x=2,z)dz + \int_{0}^{\frac{c}{c+1}} p(x=1,z)dz + \\
&= c\int_{\frac{c}{c+1}}^{1} (1-z)dz + \int_{0}^{\frac{c}{c+1}} zdz \\
&= c\left(z - \frac{z^2}{2}\right)\Bigg|_{\frac{c}{c+1}}^{1} + \frac{z^2}{2}\Bigg|_{0}^{\frac{c}{c+1}} \\
&= \frac{c}{2} - c\left(\frac{c}{c+1} - \frac{c^2}{2(c+1)^2}\right) + \frac{c^2}{2(c+1)^2} \\
&= \frac{c}{2} - \frac{c^2}{c+1} + (1+c)\frac{c^2}{2(c+1)^2} \\
&= \frac{c}{2} - \frac{c^2}{c+1} + \frac{1}{2}\frac{c^2}{c+1} \\
&= \frac{1}{2}(c - \frac{c^2}{c+1}) \\
&= \frac{1}{2}\frac{c}{c+1}
\end{aligned}
$$

5. For which values of c (the cost of damages claims) is the newly introduced machine-based test Z preferable to the human-based test Y?

   ANSWER: We compare the expected losses of the two tests. We will prefer test Z to test Y if the expected loss is lower:

$$
\begin{aligned}
\mathbb{E}_Z[L] \leq \mathbb{E}_Y[L] &\iff \frac{c}{2c+2} \leq \frac{16+c}{50} \\
&\iff 50c \leq (2c+2)(16+c) \\
&\iff 50c \leq 2c^2 + 34c + 32 \\
&\iff 0 \leq c^2 - 8c + 16 \\
&\iff 0 \leq (c-4)^2.
\end{aligned}
$$

   Since $0 \leq (c-4)^2$ holds for any cost value c, it turns out that the machine-based test is always preferable to the human-based test. Does this mean we are obsolete?

Exercise 3

The Gaussian distribution in one dimension with mean μ and variance σ^2 is

$$
\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.
$$

The *Kullback-Leibler divergence*, also known as the *relative entropy* between the distributions $p(x)$ and $q(x)$, is a measure of how much one probability distribution diverges from another. The Kullback-Leibler divergence represents the additional amount of information required to transmit the random variable x if for the coding scheme we use the "approximate" probability distribution

q, rather than the "true" distribution p. It is defined as:

$$KL(p(x)||q(x)) = -\int p(x) \ln q(x) dx + \int p(x) \ln p(x) dx. \tag{1}$$

Compute the Kullback-Leibler divergence $KL(p||q)$ between two Gaussians with the *same* variance σ^2, but different means μ and m. So $p(x) = \mathcal{N}(x|\mu, \sigma^2)$ and $q(x) = \mathcal{N}(x|m, \sigma^2)$. Verify that $KL(p||q) \geq 0$ and equal if and only if $\mu = m$.

ANSWER: Note that $\ln(p(x)) = \frac{-(x-\mu)^2}{2\sigma^2} + const(\sigma^2)$ and $\ln(q(x)) = \frac{-(x-m)^2}{2\sigma^2} + const(\sigma^2)$, where $const(\sigma^2)$ depends not on μ or m, and therefore cancels in $\ln p - \ln q$. Furthermore note that $\int p(x)xdx = \mu$ and $\int p(x)dx = 1$. Starting from Equation (1), we then have:

$$\begin{aligned}
KL(p||q) &= -\int p(x)\left\{\ln(q(x)) - \ln(p(x))\right\} dx \\
&= -\int p(x)\left\{-\frac{(x-m)^2 - (x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \int p(x)\left\{\frac{x^2 - 2mx + m^2 - x^2 + 2\mu x - \mu^2}{2\sigma^2}\right\} dx \\
&= \int p(x)\left\{\frac{-2mx + m^2 + 2\mu x - \mu^2}{2\sigma^2}\right\} dx \\
&= -\frac{2m}{2\sigma^2}\int p(x)xdx + \frac{m^2}{2\sigma^2}\int p(x)dx + \frac{2\mu}{2\sigma^2}\int p(x)xdx - \frac{\mu^2}{2\sigma^2}\int p(x)dx \\
&= -\frac{2m}{2\sigma^2}[\mu] + \frac{m^2}{2\sigma^2}[1] + \frac{2\mu}{2\sigma^2}[\mu] - \frac{\mu^2}{2\sigma^2}[1] \\
&= \frac{m^2 - 2m\mu + \mu^2}{2\sigma^2} \\
&= \frac{(m-\mu)^2}{2\sigma^2}
\end{aligned}$$

which is always greater or equal to zero since $(m - \mu)^2 \geq 0$, and obviously only equal to zero if $\mu = m$.

Exercise 4

If a random variable x has distribution $p(x)$, its entropy is

$$H[p(x)] = -\int p(x) \log p(x) dx \tag{2}$$

If two random variables x, y have joint distribution $p(x, y)$, then their entropy is defined as

$$H[p(x,y)] = -\iint p(x,y) \log p(x,y) dxdy \tag{3}$$

Use this to show that:

$$p(x,y) = p(x)p(y) \quad \Rightarrow \quad H[p(x,y)] = H[p(x)] + H[p(y)]$$

ANSWER:

$$
\begin{aligned}
H[p(x,y)] &= -\iint p(x,y)\log p(x,y)dxdy \\
&= -\iint p(x)p(y)\log\big(p(x)p(y)\big)dxdy \\
&= -\iint p(x)p(y)\big(\log p(x)+\log p(y)\big)dxdy \\
&= -\iint p(x)p(y)\log p(x)dxdy - \iint p(x)p(y)\log p(y)dxdy \\
&= -\int p(x)\log p(x)\left(\int p(y)dy\right)dx - \int p(y)\log p(y)\left(\int p(x)dx\right)dy \\
&= -\int p(x)\log p(x)dx - \int p(y)\log p(y)dy \\
&= H[p(x)] + H[p(y)]
\end{aligned}
$$

Exercise 5

Minimize $f(x,y) = 3x^2 + xy + y^2$ under constraint $x + 2y = 3$.

ANSWER: The extrema can be found by looking at the stationary points of the corresponding Lagrangian

$$L(x,y,\lambda) = 3x^2 + xy + y^2 + \lambda(x + 2y - 3)$$

Taking the gradient ∇L (partial derivatives with respect to x, y and λ) and setting equal to zero gives

$$
\begin{aligned}
\frac{\partial L}{\partial x} &= 6x + y + \lambda = 0 \\
\frac{\partial L}{\partial y} &= x + 2y + 2\lambda = 0 \\
\frac{\partial L}{\partial \lambda} &= x + 2y - 3 = 0
\end{aligned}
$$

(Note that the derivative(s) w.r.t. multipliers λ always just gives back the original constraint(s), so this step is usually implicit).
Eliminating λ and y from the first two equations yields $x = 0$. Filling in the constraints then gives $y = 1\frac{1}{2}$.

BONUS PRACTICE

Exercise 6

(Bishop 1.22) Given a loss matrix with elements L_{kj}, the expected risk is minimized if, for each $\mathbf{x}$, we choose the class that minimizes:

$$\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) \tag{4}$$

Verify that, when the loss matrix is given by $L_{kj} = 1 - \delta_{kj}$, where δ_{kj} is the Kronecker delta function, this reduces to the criterion of choosing the class having the largest posterior probability. What is the interpretation of this form of loss matrix?

ANSWER: We substitute $L_{kj} = 1 - \delta_{kj}$ into Equation 4 and we use the fact that the posterior probabilities sum to one:

$$\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) = \sum_k (1 - \delta_{kj}) p(\mathcal{C}_k|\mathbf{x}) = \sum_k p(\mathcal{C}_k|\mathbf{x}) - \sum_k \delta_{kj} p(\mathcal{C}_k|\mathbf{x}) = 1 - p(\mathcal{C}_j|\mathbf{x})$$

We find that for each $\mathbf{x}$ we should choose the class j for which $1 - p(\mathcal{C}_j|\mathbf{x})$ is a minimum, which is equivalent to choosing the j for which the posterior probability $p(\mathcal{C}_j|\mathbf{x})$ is a maximum.

Interpretation: This loss matrix assigns a loss of one if the example is misclassified, and a loss of zero if it is correctly classified, and hence minimizing the expected loss will minimize the misclassification rate.

Exercise 7

For a single binary random variable $x \in \{0, 1\}$, with $p(x = 1|\mu) = \mu$, the probability distribution over x is known as the Bernoulli distribution

$$p(x|\mu) = \mu^x (1 - \mu)^{1-x} \tag{5}$$

1. Show that this distribution satisfies the usual normalization constraint for probabilities, and compute its mean and variance.

   ANSWER: For the normalization constraint we find

   $$\sum_{x \in \{0,1\}} p(x|\mu) = p(x = 0|\mu) + p(x = 1|\mu) = (1 - \mu) + \mu = 1$$

   The mean (expectation value) is given by

   $$\sum_{x \in \{0,1\}} x p(x|\mu) = 0 \cdot p(x = 0|\mu) + 1 \cdot p(x = 1|\mu) = \mu$$

   The variance is defined as the expected squared deviation from the mean

   $$\begin{aligned} \sum_{x \in \{0,1\}} (x - \mu)^2 p(x|\mu) &= \mu^2 p(x = 0|\mu) + (1 - \mu)^2 p(x = 1|\mu) \\ &= \mu^2 (1 - \mu) + (1 - \mu)^2 \mu \\ &= \mu(1 - \mu) \end{aligned}$$

For a Bernoulli distributed variable, the loglikelihood function L as function of μ (with $0 \leq \mu \leq 1$) is given by

$$L(\mu) = \ln p(D|\mu) = m \ln \mu + (N - m) \ln(1 - \mu) \tag{6}$$

in which $m = \sum_n x_n$.

2. Assuming $0 < m < N$, show that the maximum likelihood solution is given by

$$\mu_{ML} = \frac{m}{N}$$

What do the cases $m = 0$ and $m = N$ represent? Can the solution be extended to cover these as well?

ANSWER: Differentiate (6) with respect to μ and set equal to zero

$$\frac{m}{\mu} - \frac{N - m}{1 - \mu} = 0$$

If $0 < m < N$, we see that both $\mu = 0$ and $\mu = 1$ yield a loglikelihood of minus infinity. So these points are both clearly not maxima, and we will exclude these.
Make denominator equal

$$\frac{m(1 - \mu)}{\mu(1 - \mu)} - \frac{(N - m)\mu}{\mu(1 - \mu)} = 0$$

Multiply left and right hand side with $\mu(1 - \mu)$ (excluding $\mu = 0$ and $\mu = 1$),

$$m(1 - \mu) - (N - m)\mu = 0$$

Collect terms with μ:

$$\begin{aligned} 0 &= m(1 - \mu) - (N - m)\mu \\ &= m - m\mu - N\mu + m\mu \\ &= m - N\mu \end{aligned}$$

From which the solution for the maximum likelihood follows.

Now, if $m = 0$ (only zeros), then $L = N \ln(1 - \mu)$. This is a monotonically decreasing function, so the maximum is with minimum μ, which is $\mu = 0$. With $m = 0$, this is equal to m/N.
If $m = N$ (only ones), then $L = N \ln(\mu)$. This is a monotonically increasing function, so the maximum is with maximum μ, which is $\mu = 1$. With $m = N$, this is equal to m/N.

For a discrete, binary random variable x, the entropy is given by

$$H[x] = -\sum_{x \in \{0,1\}} p(x|\mu) \log p(x|\mu) \tag{7}$$

3. Calculate the entropy (in bits) of a throw with a rather bent coin for which $p(\text{heads}) = 2/3$, and compare with a fair coin. ($\log_2(3) \approx 1.6$)

ANSWER: From eqs.(5) and (7), for Bernoulli distributed variable x we have

$$\begin{aligned} H[x] &= -\sum_{x \in \{0,1\}} p(x|\mu) \log p(x|\mu) \\ &= -\sum_{x \in \{0,1\}} \mu^x (1 - \mu)^{1-x} \{x \log(\mu) + (1 - x) \log(1 - \mu)\} \\ &= -(1 - \mu) \log(1 - \mu) - \mu \log(\mu) \end{aligned}$$

For the entropy in bits we need the $\log_2$. Using the approximation given, we have $\log_2(2/3) = \log_2(2) - \log_2(3) \approx 1 - 1.6 = -0.6$ and $\log_2(1/3) = \log_2(1) - \log_2(3) \approx 0 - 1.6 = -1.6$. Substituting in the equation above then gives for the entropy of the bent coin

$$H[x]_{bent} \approx -1/3 \cdot (-1.6) - 2/3 \cdot (-0.6) = 16/30 + 12/30 = 28/30 \approx 0.93$$

(Exact value $H[x]_{bent} = 0.9183...$). For the fair coin we simply have $\log_2(1/2) = -1$, so

$$H[x]_{fair} = -1/2 \cdot (-1) - 1/2 \cdot (-1) = 1/2 + 1/2 = 1$$

So we find (as to be expected) that the fair coin has a higher entropy than the bent coin, reflecting the fact that a fair coin is 'maximally unpredictable'.

The form of the Bernoulli distribution is not symmetric between the two values of x. Sometimes, it is more convenient to use an equivalent formulation for which $x \in \{-1, 1\}$. The binary distribution over x can then be written in an exponential form

$$p(x|\theta) = \frac{1}{Z(\theta)} \exp(x\theta) \tag{8}$$

with parameter $-\infty < \theta < \infty$.

4. Compute $Z(\theta)$. What is roughly the chance on $x = -1$ when $\theta \approx 1$?

ANSWER: $Z(\theta)$ is the normalizing constant that depends on the value for θ. By just filling in $x = -1$ and $x = 1$ we see that the probability of the two possible outcomes is

$$\begin{aligned} p(x = -1|\theta) &= \frac{1}{Z(\theta)} \exp(-\theta) \\ p(x = 1|\theta) &= \frac{1}{Z(\theta)} \exp(\theta) \end{aligned}$$

Since $p(x = -1|\theta) + p(x = 1|\theta) = 1$, and filling in, we can conclude that

$$Z(\theta) = \exp(-\theta) + \exp(\theta)$$

Substituting in (8), with $Z(1) = \exp^{-1} + \exp$ and $\exp \approx 2.718282$, gives roughly

$$p(x = -1|\theta = 1) = \frac{1}{Z(1)} \frac{1}{\exp} \approx \left(\frac{1}{> 3}\right) \left(\frac{1}{< 3}\right) \approx \frac{1}{9}$$

(Actual value = 0.1192...).

Exercise 8

A factory produces products X. 20% is of quality $x = 1$ and the remainder of quality $x = 2$. There is a test Z, which can have an outcome $\{1, 2, 3, 4, 5\}$. The conditional probability density of z, depending on the quality x is

$p(z = 1|x = 1) = 0.15$; $p(z = 2|x = 1) = 0.15$; $p(z = 3|x = 1) = 0.4$; $p(z = 4|x = 1) = 0.25$; $p(z = 5|x = 1) = 0.05$
$p(z = 1|x = 2) = 0.12$; $p(z = 2|x = 2) = 0.18$; $p(z = 3|x = 2) = 0.2$; $p(z = 4|x = 2) = 0.22$; $p(z = 5|x = 2) = 0.28$

Suppose we observe test result $z = 3$. Compute, using Bayes' rule, the posterior probability $p(x = 1|z = 3)$.

ANSWER:

$$p(x = 1|z = 3) = \frac{p(z = 3|x = 1)p(x = 1)}{p(z = 3|x = 1)p(x = 1) + p(z = 3|x = 2)p(x = 2)} = \frac{0.4 \times 0.2}{0.4 \times 0.2 + 0.2 \times 0.8} = \frac{1}{3}$$

and $p(x = 2|z = 3) = 1 - p(x = 1|z = 3) = 2/3$