

Machine Learning: Exercises for Block 1 (Lectures 1-4)

Isabel Valera

Exercise 1: Fruits

Suppose that we have three coloured boxes r (red), b (blue), and g (green). Box r contains 3 apples, 4 oranges, and 3 limes, box b contains 1 apple, 1 orange, and 0 limes, and box g contains 3 apples, 3 oranges, and 4 limes. A box is chosen at random with probabilities $p(r) = 0.2, p(b) = 0.2, p(g) = 0.6$, and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the selected box).

- i) What is the probability of selecting an apple?
- ii) If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

Exercise 2: Maximum Density

Consider a probability density $p_x(x)$ defined over a continuous variable x , and suppose that we make a nonlinear change of variable using $x = g(y)$, so that the density transforms according to

$$p_y(y) = p_x(g(y))|g'(y)| \quad (1)$$

- i) By differentiating [1](#), show that the location \hat{y} of the maximum of the density in y is not in general related to the location \hat{x} of the maximum of the density over x by the simple functional relation $\hat{x} = g(\hat{y})$ as a consequence of the Jacobian factor. This shows that the maximum of a probability density (in contrast to a simple function) is dependent on the choice of variable.
- ii) Verify that, in the case of a linear transformation, the location of the maximum transforms in the same way as the variable itself.

Exercise 3: Variance

Let $f(x)$ be some function in x . Using the definition $\text{var}[f] = \mathbb{E} \left[(f(x) - \mathbb{E}[f(x)])^2 \right]$ (c.f. [\[1\] 1.38](#)) of the variance show that $\text{var}[f(x)]$ satisfies $\text{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$

Exercise 4: Covariance

Show that if two variables x and y are independent, then their covariance is zero.

Exercise 5: Normal Mode

Recall the definition of the univariate Gaussian distribution

$$\text{Gauss}(\mathbf{x}|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (2)$$

and the definition of the multivariate D -dimensional Gaussian distribution

$$\text{Gauss}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \quad (3)$$

- i) Show that the mode (i.e. the maximum) of the Gaussian distribution 2 is given by μ .
- ii) Show that the mode of the multivariate Gaussian 3 is given by $\boldsymbol{\mu}$.

Exercise 6: Independence

Suppose that the two variables x and z are statistically independent.

- i) Show that the mean satisfies $\mathbb{E}[x + z] = \mathbb{E}[x] + \mathbb{E}[z]$.
- ii) Show that the variance satisfies $\text{var}[x + z] = \text{var}[x] + \text{var}[z]$.

Exercise 7: Maximum likelihood estimates

Verify by setting the derivatives of the log likelihood

$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) \quad (4)$$

with respect to μ and σ^2 equal to zero:

- i) $\mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n$ (see [1] 1.55)
- ii) $\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2$ (see [1] 1.56)

Exercise 8: True variance

Suppose that the variance of a Gaussian is estimated using $\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2$ (see [1] 1.56) but with the maximum likelihood estimate μ_{ML} replaced with the true value μ of the mean. Show that this estimator has the property that its expectation is given by the true variance σ^2

Exercise 9: Symmetry

Show that an arbitrary square matrix with elements w_{ij} can be written in the form $w_{ij} = w_{ij}^S + w_{ij}^A$ where w_{ij}^S and w_{ij}^A are symmetric and anti-symmetric matrices, respectively, satisfying $w_{ij}^S = w_{ji}^S$ and $w_{ij}^A = -w_{ji}^A$ for all i and j . Now consider the second order term in a higher order polynomial in D dimensions given by

$$\sum_{i=1}^D \sum_{j=1}^D w_{ij} x_i x_j$$

i) Show that

$$\sum_{i=1}^D \sum_{j=1}^D w_{ij} x_i x_j = \sum_{i=1}^D \sum_{j=1}^D w_{ij}^S x_i x_j$$

so that the contribution from the anti-symmetric matrix vanishes.

ii) We see that, without loss of generality, the matrix of coefficients w_{ij} can be chosen to be symmetric, and so not all of the D^2 elements of this matrix can be chosen independently. Show that the number of independent parameters in the matrix w_{ij}^S is given by $D(D+1)/2$.

Exercise 10: Misclassification bound

Consider two nonnegative numbers a and b , and show that, if $a \leq b$, then $a \leq \sqrt{ab}$. Use this result to show that, if the decision regions of a two-class classification problem with classes $\mathcal{C}_1, \mathcal{C}_2$ are chosen to minimize the probability of misclassification, this probability will satisfy

$$p(\text{mistake}) \leq \int \{p(x, \mathcal{C}_1)p(x, \mathcal{C}_2)\}^{1/2} dx$$

Exercise 11: Minimal loss (i)

Given a loss matrix with elements L_{kj} , the expected risk is minimized if, for each x , we choose the class that minimizes

$$\sum_k L_{kj} p(\mathcal{C}_k | x) \quad (5)$$

- i) Verify that, when the loss matrix is given by $L_{kj} = 1 - I_{kj}$ where I_{kj} are the elements of the identity matrix, this reduces to the criterion of choosing the class having the largest posterior probability.
- ii) What is the interpretation of this form of loss matrix?

Exercise 12: Minimal loss (ii)

Derive the criterion for minimizing the expected loss when there is a general loss matrix and general prior probabilities for the classes.

Exercise 13: Targets

Consider the generalization of the squared loss function

$$L(t, y(x)) = \{y(x) - t\}^2 \quad (6)$$

for a single target variable t to the case of multiple target variables described by the vector \mathbf{t} given by

$$\mathbb{E}[L(\mathbf{t}, \mathbf{y}(\mathbf{x}))] = \int \int \|\mathbf{y}(\mathbf{x}) - \mathbf{t}\|^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t}$$

- i) Using the **calculus of variations**, show that the function $\mathbf{y}(\mathbf{x})$ for which this expected loss is minimized is given by $\mathbf{y}(\mathbf{x}) = \mathbb{E}_{\mathbf{t}}[\mathbf{t}|\mathbf{x}]$.
- ii) Show that this result reduces to $y(x) = \mathbb{E}_t[t|\mathbf{x}]$ (c.f. [1] (1.89)) for the case of a single target variable t .

Exercise 14: Regression

Consider the expected loss for regression problems under the L_q loss function given by

$$\mathbb{E}[L_q] = \int \int |y(\mathbf{x}) - t|^q p(\mathbf{x}, t) d\mathbf{x} dt \quad (7)$$

- i) Write down the condition that $y(\mathbf{x})$ must satisfy in order to minimize $\mathbb{E}[L_q]$.
- ii) Show that, for $q = 1$, this solution represents the conditional median, i.e., the function $y(\mathbf{x})$ such that the probability mass for $t < y(\mathbf{x})$ is the same as for $t \geq y(\mathbf{x})$.
- iii) Show that the minimum expected L_q loss for $q \rightarrow 0$ is given by the conditional mode, i.e., by the function $y(\mathbf{x})$ equal to the value of t that maximizes $p(t|\mathbf{x})$ for each \mathbf{x}

Exercise 15: Decision boundary

Consider the following decision rule for a two-category one-dimensional problem: Decide \mathcal{C}_1 if $x > \theta$; otherwise decide \mathcal{C}_2 .

- i) Show that the probability of error for this rule is given by

$$P(\text{error}) = P(\mathcal{C}_1) \int_{-\infty}^{\theta} p(x|\mathcal{C}_1) dx + P(\mathcal{C}_2) \int_{\theta}^{\infty} p(x|\mathcal{C}_2) dx$$

- ii) By differentiating, show that a necessary condition to minimize $P(\text{error})$ is that θ satisfy

$$p(\theta|\mathcal{C}_1)P(\mathcal{C}_1) = p(\theta|\mathcal{C}_2)P(\mathcal{C}_2) \quad (8)$$

- iii) Does equation 8 define θ uniquely?
- iv) Give an example where a value of θ satisfying the equation actually maximizes the probability of error.

Exercise 16: At the limit

Let $\mathbf{x} = (x_1, \dots, x_d)^T$ be binary valued and $P(\mathcal{C}_j)$ be the prior probability for class \mathcal{C}_j and $j \in \{1, 2\}$. Now define

$$\begin{aligned}P(x_i = 1|\mathcal{C}_1) &= p_{i1} = p > \frac{1}{2} \\P(x_i = 1|\mathcal{C}_2) &= p_{i2} = 1 - p \\P(\mathcal{C}_1) &= P(\mathcal{C}_2) = \frac{1}{2}\end{aligned}$$

with $i \in \{1, \dots, d\}$ and d odd.

i) Show that the minimum-error-rate decision rule becomes:

$$\text{Decide } \mathcal{C}_1 \text{ if } \sum_{i=1}^d x_i > \frac{d}{2} \text{ and } \mathcal{C}_2 \text{ otherwise}$$

ii) Show that the minimum probability of error is given by

$$P_e(d, p) = \sum_{k=0}^{(d-1)/2} \binom{d}{k} p^k (1-p)^{d-k}$$

where $\binom{d}{k} = \frac{d!}{k!(d-k)!}$ is the binomial coefficient.

iii) What is the limiting value of $P_e(d, p)$ as $p \rightarrow \frac{1}{2}$? Explain.

iv) Show that $P_e(d, p)$ approaches zero as $d \rightarrow \infty$. Explain.

References

- [1] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.
- [2] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification. A Wiley-Interscience Publication*, 2001.