Machine Learning 2024 - Sheet 1

Isabel Valera

Exercise 1: Fruits ⚡

Suppose that we have three colored 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?

Note: The solutions are $p(\text{apple}) = 0.34$ and $p(\text{green}|\text{orange}) = 0.5$.

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

$$p_y(y) = p_x(g(y)) * |g'(y)| \tag{1}$$

i) By differentiating the above equation, show that the location $\hat{y}$ of the maximum (i.e. the mode) 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 $var[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]$ of the variance show that $var[f(x)]$ satisfies $var[f] = \mathbb{E}\left[f(x)^2\right] - \mathbb{E}[f(x)]^2$

Exercise 4: Normal Mode ⚡⚡

Recall the definition of the univariate Gaussian distribution

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

and the definition of the multivariate D-dimensional Gaussian distribution

$$\mathcal{N}\left(x|\boldsymbol{\mu},\Sigma\right)=\frac{1}{\sqrt{(2\pi)^D|\Sigma|}}\exp\left(-\frac{1}{2}(x-\boldsymbol{\mu})^T\Sigma^{-1}(x-\boldsymbol{\mu})\right) \tag{3}$$

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 5: Maximum likelihood estimates ⚡⚡

You are given a dataset $X=\{x_i\}_{i=1}^N$ of i.i.d. samples from a gaussian distribution with unknown mean and variance.

Verify, by setting the derivatives of the log likelihood

$$\ln p(X|\mu,\sigma^2)=-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2-\frac{N}{2}\ln\sigma^2-\frac{N}{2}\ln(2\pi) \tag{4}$$

with respect to μ (and with respect to σ^2) equal to zero, that the maximum likelihood estimates for the mean and variance of the true underlying gaussian distribution are given by:

i) $\mu_{ML}=\frac{1}{N}\sum_{n=1}^{N}x_n$

ii) $\sigma^2_{ML}=\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$

Exercise 6: Misclassification bound ⚡⚡⚡

Consider two nonnegative numbers a and b, and show that, if $a\leq b$, then $a\leq\sqrt{ab}$. Consider now a two-class classification problem (i.e., $\mathcal{Y}=\{-1,1\}$). Use the above result to show that, if the decision regions (i.e. the two regions on $\mathcal{X}$ where we decide for $Y=-1$ and $Y=+1$, respectively) are chosen to minimize the probability of misclassification, this probability will satisfy

$$p(\text{mistake})\leq\int\sqrt{(p(x,Y=-1)p(x,Y=1)}dx$$

Hint: You will probably find it useful to give the decision regions names, i.e. R_{-1} and R_{+1}, and express the probability of mistake as

$$p(\text{mistake})=p\left(x\in\mathcal{R}_{-1},Y=+1\right)+p\left(x\in\mathcal{R}_{+1},Y=-1\right)=\int_{\mathcal{R}_{-1}}p\left(x,Y=+1\right)\mathrm{d}x+\int_{\mathcal{R}_{+1}}p\left(x,Y=-1\right)\mathrm{d}x$$

Exercise 7: Minimal loss ⚡

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

$$\sum_k C_{kj} p(Y = k|x) \tag{5}$$

i) Verify that, when the cost matrix is given by $C_{kj} = 1 - I_{kj}$ 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 cost matrix?

Hint: Write down the cost matrix explicitly!

Exercise 8: Cost-sensitive Bayes classification ⚡⚡

Consider cost-sensitive Bayes classification with two classes. The class conditionals are gaussian distributions:

$$p(Y = -1) = \frac{1}{4} \quad \text{and} \quad p(Y = +1) = \frac{3}{4}$$

$$p(x|Y = -1) = \frac{1}{\sqrt{2\pi}} \exp^{[-\frac{1}{2}(x-2)^2]}$$

$$p(x|Y = +1) = \frac{1}{\sqrt{2\pi}} \exp^{[-\frac{1}{2}(x-5)^2]}$$

The costs are given by a two-by-two cost matrix C_{ij}:

$$C_{ij} = C(Y = i, \hat{y}(X) = j) = \begin{pmatrix} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{pmatrix}$$

You can assume that $\lambda_{12} > 0$ and $\lambda_{21} > 0$.

i) Derive the posterior probabilities $p(Y = -1|x)$ and $p(Y = +1|x)$.

ii) Write down the expected loss for each of the two decisions.

iii) Derive the cost-sensitive Bayes classifier explicitly, that means give a condition that determines when class 1 should be chosen.

Exercise 9: More Bayesian decision theory ⚡

Consider now a binary classification problem $\mathcal{Y} = \{-1, 1\}$, with the following distribution on $\mathcal{X} = [0, 1]$,

$$\mathrm{P}(Y = 1 \,|\, X = x) = \begin{cases} 0.2, & \text{if } 0 \leq x \leq 0.25 \\ 0.8, & \text{if } 0.25 < x < 0.75 \\ 0.2, & \text{if } 0.75 \leq x \leq 1 \end{cases}$$

where x is uniformly sampled from $[0, 1]$.

i) What is the Bayes optimal error of this problem?

ii) Report the optimal set of parameters (w^*, b^*) (i.e., those that minimize the error probability), and the resulting error probability for a classifier of the form

$$f_{(w,b)} = sign(wx + b), \qquad w, b \in \mathbb{R}.$$

Hint: The set of optimal parameters is best expressed as a union of two of its subsets. Write the subsets in set notation and use the $\cup$ notation to combine them.

Correct end result: 0.2 for *i*) and 0.35 for *ii*).

Exercise 10: Decision boundary ⚡⚡⚡

Consider the following decision rule for a two-class one-dimensional problem: Decide for $Y = -1$ if $x > \theta$; otherwise decide for $Y = +1$.

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

$$P(error) = P(Y = -1) \int_{-\infty}^{\theta} p(x|Y = -1) + P(Y = +1) \int_{\theta}^{\infty} p(x|Y = +1)dx$$

ii) By differentiating, show that a necessary condition to minimize $P(error)$ is that θ satisfy

$$p(\theta|Y = -1)P(Y = -1) = p(\theta|Y = +1)P(Y = +1) \tag{6}$$

iii) Does equation 6 define θ uniquely?

iv) Give an example where a value of θ satisfying the equation actually maximizes the probability of error.

Exercise 11: Targets ⚡⚡⚡

Consider the generalization of the squared loss function

$$L(y, f(x)) = (f(x) - y)^2 \tag{7}$$

for a single target variable y to the case of multiple target variables described by the vector y given by

$$\mathbb{E}\left[L(y, f(x))\right] = \int \int ||f(x) - y||^2 p(x, y)dxdy$$

Show that the function $f(x)$ for which this expected loss is minimized is given by $f(x) = \mathbb{E}_y\left[y|x\right]$. To do this, derive $\frac{\delta \mathbb{E}[L]}{\delta f(x)}$, set it to zero, and solve for $f(x)$.

Exercise 12: Regression ⚡⚡

(i) Ridge regression with some twists: Given a sample $D_n = (X_i, Y_i)_{i=1}^n$, ridge regression is formulated as:

$$D_n \mapsto w_n = \arg\min_{w \in \mathcal{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \langle \mathbf{w}, \Phi\left(X_i\right)\rangle\right)^2 + \lambda \|\mathbf{w}\|_2^2, \lambda > 0 \tag{8}$$

(a) What if instead of $+\lambda\|\mathbf{w}\|_2^2$, we accidentally wrote $+\lambda * Y^T Y$ instead? Explain the effect this "regularization" would have.

(b) Consider the original formulation in 8 again, but we choose a $\lambda < 0$ now. What effect does the regularization now have? Does it still serve the purpose of regularization?

(c) True or false? Say you use the original ridge regression formulation as in 8. If all features x are rescaled by a constant term and you re-run ridge regression, the test set prediction accuracy will change. In other words, ridge regression test set performance varies with the scale of the datasets features.

(ii) Suppose you have a linear-regression model in one dimension, with β_0 (intercept) equal to zero, and β_1 equal to 2, and you are using mean-squared error without regularization. Given the input-output pairs $\{(10, 22), (1, 2.5), (2, 3)\}$, compute the gradient of the error w.r.t. β_1.

Exercise 13: More regression

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

$$\mathbb{E}\left[L_q\right] = \int\int |f(x) - y|^q p(x, y) dx dy \tag{9}$$

i) Write down the condition that $f(x)$ must satisfy in order to minimize $\mathbb{E}\left[L_q\right]$.

ii) Show that, for $q = 1$, this solution represents the conditional median, i.e., the function $f(x)$ such that the probability mass for $y < f(x)$ is the same as for $y \geq f(x)$.

iii) Show that the minimum expected L_q loss for $q \to 0$ is given by the conditional mode, i.e., by the function $f(x)$ equal to the value of y that maximizes $p(y|x)$ for each x