
Exercise 1: Bayes' rule (6 points). Suppose that 0.1% of all credit card transactions are fraudulent. And suppose that there is a deployed ML model to automatically detect fraud which has a 0.2% false positive rate and a 0.5% false negative rate ("positive" refers to the fraudulent class).

1. (3 points) The ML model labels Transaction A as positive. What is the probability that transaction A is actually fraudulent?
2. (3 points) The ML model labels Transaction B as negative. What is the probability that transaction B is actually valid (non-fraudulent)?

Exercise 2: Bayesian decision theory: losses and risks (11 points). Consider a classification problem with K classes, using a loss $\lambda_{ik} \geq 0$ if we choose class i when the input actually belongs to class k, for $i, k \in \{1, \dots, K\}$.

1. (2 points) Write the expression for the expected risk $R_i(\mathbf{x})$ for choosing class i as the class for a pattern $\mathbf{x}$, and the rule for choosing the class for $\mathbf{x}$.

Consider a two-class problem with losses given by the matrix $\lambda_{ik} = \left(\begin{smallmatrix} 0 & 1 \\ \lambda_{21} & 0 \end{smallmatrix}\right)$.

2. (3 points) Give the optimal decision rule in the form "$p(C_1|\mathbf{x}) > \dots$" as a function of λ_{21}.
3. (3 points) Imagine we consider both misclassification errors as equally costly. When is class 1 chosen (for what values of $p(C_1|\mathbf{x})$)?
4. (3 points) Imagine we want to be very conservative when choosing *class 2* and we seek a rule of the form "$p(C_2|\mathbf{x}) > 0.9$" (i.e., choose class 2 when its posterior probability exceeds 90%). What should λ_{21} be?

Exercise 3: association rules (6 points). Given the following data of transactions at a supermarket, calculate the support and confidence values of the following association rules: beer → diapers, diapers → beer, beer → milk, milk → beer, milk → diapers, diapers → milk. What is the best rule to use in practice?

transaction #	items in basket
1	milk, diapers
2	milk
3	beer, milk
4	beer, milk, diapers
5	beer
6	milk, diapers

Exercise 4: true- and false-positive rates (10 points). We have a dataset with $N = 5$ points for binary classification as given by the following table, where $\mathbf{x}_n$ is a pattern, y_n its ground-truth label (1 = positive class, 2 = negative class) and $p(C_1|\mathbf{x}_n)$ the posterior probability produced by some probabilistic classification algorithm:

n	1	2	3	4	5
y_n	1	2	2	1	1
$p(C_1\|\mathbf{x}_n)$	0.9	0.2	0.7	0.5	0.4

We use a classification rule of the form "$p(C_1|\mathbf{x}) > \theta$" where $\theta \in [0, 1]$ is a threshold.

1. (8 points) Give, *for all possible values of* $\theta \in [0, 1]$, the predicted labels and the corresponding confusion matrix and classification error.
2. (2 points) Plot the corresponding pairs (fp, tp) as an ROC curve.

Exercise 5: least-squares regression (14 points). Consider the following model, with parameters $\mathbf{\Theta} = \{\alpha_1, \alpha_2, \alpha_3\} \subset \mathbb{R}$ and an input $x \in \mathbb{R}$:

$$h(x; \alpha_1, \alpha_2, \alpha_3) = \alpha_1 + \alpha_2 x + \alpha_3 e^{-x} \in \mathbb{R}.$$

1. (2 points) Write the general expression of the least-squares error function of a model $h(x; \mathbf{\Theta})$ with parameters $\mathbf{\Theta}$ given a sample $\{(x_n, y_n)\}_{n=1}^N$.
2. (2 points) Apply it to the above model, simplifying it as much as possible.
3. (6 points) Find the least-squares estimate for the parameters.
4. (4 points) Assume the values $\{x_n\}_{n=1}^N$ are uniformly distributed in the interval $[0, 2\pi]$. Can you find a simpler, approximate way to find the least-squares estimate ? *Hint*: approximate $\frac{1}{N}\sum_{n=1}^N f(x_n)$ by an integral.

Exercise 6: maximum likelihood estimate (15 points). Consider a real random variable $x \in \mathbb{R}$ which the following probability density function:

$$p(x; \sigma) = \frac{x}{\sigma^2}\, e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} \qquad x \geq 0$$

where the parameter is $\sigma > 0$.

1. (2 points) Verify that $\int_0^\infty p(x)\, dx = 1$.
2. (2 points) Write the general expression of the log-likelihood of a density $p(x; \mathbf{\Theta})$ with parameters $\mathbf{\Theta}$ for an iid sample $x_1, \dots, x_N \in \mathbb{R}$.
3. (5 points) Apply it to the above distribution, simplifying it as much as possible.
4. (6 points) Find the maximum likelihood estimate for the parameters.

Exercise 7: exponential classifiers (18 points). We have a binary classification problem on an input $x \geq 0$ where the distribution of class $k \in \{1, 2\}$ is exponential (with parameter $\lambda_k \geq 0$):

$$p(x|C_k) = \lambda_k\, e^{-\lambda_k x} \qquad x \geq 0$$

and each class has a prior probability $p(C_k)$. All the parameters $\{p(C_k), \lambda_k\}_{k=1}^2$ have been fixed in a previous training step. Assume $\lambda_2 > \lambda_1$.

1. (2 points) Verify that $\int_0^\infty p(x|C_k)\, dx = 1$.
2. (2 points) Write the general expression for the posterior distribution $p(C_k|\mathbf{x})$ (using Bayes' theorem).
3. (5 points) Apply it to our case and simplify the result as much as possible.
4. (4 points) Consider the usual classification rule "predict class 1 if $p(C_1|x) > \frac{1}{2}$". Apply it to our case and determine the region of input space ($x \geq 0$) that belongs to each class, and the class boundaries.
5. (5 points) Define class discriminant functions $g_k(x) = \log p(x|C_k) + \log p(C_k)$ for $k \in \{1, 2\}$, where the class predicted for x is $\arg\max_{k=1,\dots,K} g_k(x)$. Verify they produce the same class boundaries.

Exercise 8: multivariate Bernoulli distribution (20 points). Consider a multivariate Bernoulli distribution where $\boldsymbol{\theta} \in [0, 1]^D$ are the parameters and $\mathbf{x} \in \{0, 1\}^D$ the binary random vector:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{d=1}^D \theta_d^{x_d}\, (1 - \theta_d)^{1 - x_d}.$$

1. (5 points) Compute the maximum likelihood estimate for $\boldsymbol{\theta}$ given a sample $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$.

Let us do document classification using a D-word dictionary (element d in $\mathbf{x}_n$ is 1 if word d is in document n and 0 otherwise) using a multivariate Bernoulli model for each class. Assume we have K document classes for which we have already obtained the values of the optimal parameters $\boldsymbol{\theta}_k = (\theta_{k1}, \dots, \theta_{kD})^T$ and prior distribution $p(C_k) = \pi_k$, for $k = 1, \dots, K$, by maximum likelihood.

2. (2 points) Write the discriminant function $g_k(\mathbf{x})$ for a probabilistic classifier in general (not necessarily Bernoulli), and the rule to make a decision.

3. (5 points) Apply it to the multivariate Bernoulli case with K classes. Show that $g_k(\mathbf{x})$ is linear on $\mathbf{x}$, i.e., it can be written as $g_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + w_{k0}$ and give the expression for $\mathbf{w}_k$ and w_{k0}.

4. (3 points) Consider $K = 2$ classes. Show the decision rule can be written as "if $\mathbf{w}^T \mathbf{x} + w_0 > 0$ then choose class 1", and give the expression for $\mathbf{w}$ and w_0.

5. (5 points) Compute the numerical values of $\mathbf{w}$ and w_0 for a two-word dictionary where $\pi_1 = 0.6$, $\boldsymbol{\theta}_1 = \left(\begin{smallmatrix} 0.4 \\ 0.1 \end{smallmatrix}\right)$ and $\boldsymbol{\theta}_2 = \left(\begin{smallmatrix} 0.3 \\ 0.5 \end{smallmatrix}\right)$. Plot in 2D all the possible values of $\mathbf{x} \in \{0, 1\}^D$ and the boundary corresponding to this classifier.