Learning Machine Learning

tommyod

October 7, 2018

Abstract

This document contains my notes, and solutions to, the book "Pattern Classification" by Duda et al.

Contents

1 Notes from "Pattern Recognition"

2 Solutions to "Pattern Recognition"

2.1 Bayesian Decision Theory

Problem 2.6

a) We have to find x^* such that $p(\alpha_2, \omega_1) \leq E_1$, i.e. the probability of choosing α_2 and the true state of nature being ω_1 is smaller than some prescribed limit E_1.

$$p(\alpha_2, \omega_1) = p(\alpha_2|\omega_1)P(\omega_1) = p(x > x^*|\omega_1)P(\omega_1)$$

Let the cumulative gaussian be given by $g(x)$, then we have

$$p(x > x^*|\omega_1)P(\omega_1) = \left(1 - g\left(x < x^*|\omega_1\right)\right)P(\omega_1) = \left(1 - g\left(x < \frac{x^* - \mu_1}{\sigma_1}\right)\right)P(\omega_1),$$

which means that

$$\left(1 - \Phi\left(\frac{x^* - \mu_1}{\sigma_1}\right)\right)P(\omega_1) \leq E_1 \quad \Rightarrow \quad x^* \geq \mu_1 + \sigma_1 \Phi^{-1}\left(1 - \frac{E_1}{P(\omega_1)}\right).$$

Two sanity checks are in order. First, notice that as $E_1 \to 0$ the argument of ϕ^{-1} goes to 1 and x^* goes to infinity. In words, this means that if we want to avoid every E_1 error we have to classify *every* observation as ω_1.

Notice also that if we choose $E_1 = 0.5$, the the argument of ϕ^{-1} becomes 0, goes to 1 and x^* goes to infinity.

Problem 2.12

a) The key observation is that the maximal value $P(\omega_{\max}|\boldsymbol{x})$ is greater than, or equal to, the average. Therefore we obtain

$$P(\omega_{\max}|\boldsymbol{x}) \geq \frac{1}{c}\sum_{i=1}^{c} P(\omega_i|\boldsymbol{x}) = \frac{1}{c},$$

where the last equality is due to probabilities summing to unity.

b) The minimum error rate is achieved by choosing $\omega_{\max}$, the most likely state of nature. The average probability of error over the data space is therefore the probability that $\omega_{\max}$ is *not* the true state of nature for a given $\boldsymbol{x}$, that is:

$$P(\text{error}) = \mathbb{E}_x \left[1 - P(\omega_{\max}|\boldsymbol{x})\right] = 1 - \int P(\omega_{\max}|\boldsymbol{x}) p(\boldsymbol{x})\, d\boldsymbol{x}.$$

c) We see that

$$P(\text{error}) = 1 - \int P(\omega_{\max}|\boldsymbol{x}) p(\boldsymbol{x})\, d\boldsymbol{x} \leq 1 - \int \frac{1}{c} p(\boldsymbol{x})\, d\boldsymbol{x} = 1 - \frac{1}{c} = \frac{c-1}{c}.$$

d) A situation where $P(\text{error}) = (c-1)/c$ arises when $P(\omega_i) = 1/c$. Then the maximum value is equal to the average value, and the inequality in part a) becomes an equality.

Problem 2.19

a) The entropy is given by $\mathrm{H}\left[p(x)\right] = -\int p(x) \ln p(x)\, dx$. The optimization problem gives the synthetic function

$$H_s = -\int p(x) \ln p(x)\, dx + \sum_{k=1}^{q} \lambda_k \left(\int b_k(x) p(x)\, dx - a_k \right),$$

and since a probability density function has $\int p(x)\, dx = 1$ we add an additional constraint for $k = 0$ with $b_0(x) = 1$ and $a_k = 1$. Collecting terms we obtain

$$\begin{aligned} H_s &= -\int p(x) \ln p(x)\, dx + \sum_{k=0}^{q} \lambda_k \int b_k(x) p(x)\, dx - \sum_{k=0}^{q} \lambda_k a_k \\ &= -\int p(x) \left[\ln p(x) - \sum_{k=0}^{q} \lambda_k b_k(x) \right] dx - \sum_{k=0}^{q} \lambda_k a_k, \end{aligned}$$

which is what we were asked to show.

b) Differentiating the equation above with respect to $p(x)$ and equating it to zero we obtain

$$-\int \left(1 \left[\ln p(x) - \sum_{k=0}^{q} \lambda_k b_k(x) \right] + p(x) \left[\frac{1}{p(x)} \right] \right) dx = 0.$$

This integral is zero if the integrand is zero for every x, so we require that

$$\ln p(x) - \sum_{k=0}^{q} \lambda_k b_k(x) + 1 = 0,$$

and solving this equation for $p(x)$ gives the desired answer.

Problem 2.21

We are asked to compute the entropy of the Gaussian, triangle distribution and uniform distribution. Every p.d.f has $\mu = 0$ and standard deviation σ.

Gaussian We use the definition $\mathrm{H}[p(x)] = -\int p(x) \ln p(x)\, dx$ to compute

$$\mathrm{H}[p(x)] = -\int \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right) \left[\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2}\frac{x^2}{\sigma^2} \right] dx.$$

Let us denote $K = \frac{1}{\sqrt{2\pi}\sigma}$ to simplify notation. We obtain

$$-\int K \exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right) \left[\ln K - \frac{1}{2}\frac{x^2}{\sigma^2} \right] dx =$$

$$-K \ln K \int \exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right) dx + K \int \frac{1}{2}\frac{x^2}{\sigma^2} \exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right) dx$$

The first term is simply $-\ln K$, since it's the normal distribution with an additional factor $-\ln K$. The second term is not as easy. We change variables to $y = x/\left(\sqrt{2}\sigma\right)$, and write it as

$$K \int y^2 \exp\left(-y^2\right) \sqrt{2}\sigma\, dy,$$

which can be solved by using the following observation (integration by parts)

$$\int 1 e^{-x^2}\, dx = \underbrace{\left. x e^{-x^2} \right|}_{\text{0 due to symmetry}} - \int x(-2x) e^{-x^2}\, dx.$$

We proceed by using this fact, and integrate as follows:

$$K\sqrt{2}\sigma \int y^2 \exp\left(-y^2\right)\, dy = K\sqrt{2}\sigma \frac{1}{2} \int \exp\left(-y^2\right)\, dy = K\sqrt{2}\sigma \frac{1}{2}\sqrt{\pi} = \frac{1}{2}$$

To recap, the first integral evaluated to $-\ln K$, and the second evaluated to $\frac{1}{2}$. The entropy of the Gaussian is $1/2 + \ln \sqrt{2\pi}\sigma$.

Triangle The triangle distribution is of the form

$$f(x) = \begin{cases} h - \frac{hx}{b} & \text{if } |x| < b \\ 0 & \text{if } |x| \geq b, \end{cases}$$

where h is a height and b is the width to the left of, and to the right of, $x = 0$.

Since the integral must evaluate to unity, we impose $hb = 1$ and obtain $f(x; b) = \frac{1}{b}\left(1 - \frac{x}{b}\right)$. We wish to parameterize the triangle distribution using the standard deviation σ instead of width b. We can use $\text{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ to find the variance, since in this case $\mathbb{E}(X)^2 = \mu^2 = 0$ since the function is centered on $x = 0$. Computing $\mathbb{E}(X^2)$ yields $b^2/6$, so $b^2 = 6\sigma^2$. The revised triangle distribution then becomes

$$f(x; \sigma) = \begin{cases} \frac{1}{\sqrt{6}\sigma}\left(1 - \frac{x}{\sqrt{6}\sigma}\right) & \text{if } |x| < \sqrt{6}\sigma \\ 0 & \text{if } |x| \geq \sqrt{6}\sigma. \end{cases}$$

We set $k = \frac{1}{\sqrt{6}\sigma}$ to ease notation. Due to symmetry, we compute the entropy as

$$\mathrm{H}\left[f(x; \sigma)\right] = -2 \int_0^{\sqrt{6}\sigma} k\,(1 - kx) \ln\left(k\,(1 - kx)\right)\, dx.$$

Changing variables to $y = 1 - kx$ we obtain

$$-2 \int_{x=0}^{x=\sqrt{6}\sigma} ky\,(\ln k + \ln y)\; dx = -2 \int_{y=1}^{y=0} ky\,(\ln k + \ln y)\left(\frac{1}{-k}\right) dy$$
$$-2 \int_0^1 y\,(\ln k + \ln y)\, dy = -2 \int_0^1 y \ln k\, dy - 2 \int_0^1 y \ln y\, dy = -2\left(\ln k - \frac{1}{4}\right),$$

where the last integral can be evaluated using integration by parts. The entropy of the triangle distribution is $1/2 + \ln \sqrt{6}\sigma$.

Uniform Using the same logic as with the triangle distribution to normalize a uniform distribution, and then parameterizing by σ, we obtain

$$u(x; \sigma) = \left\{\begin{array}{cl} \dfrac{1}{2b} & \text{if } |x| < b \\ 0 & \text{if } |x| \geq b \end{array}\right\} = \left\{\begin{array}{cl} \dfrac{1}{2\sqrt{3}\sigma} & \text{if } |x| < \sqrt{3}\sigma \\ 0 & \text{if } |x| \geq \sqrt{3}\sigma. \end{array}\right.$$

Computing the entropy is easier than in the case of the Gaussian and the triangular distribution, we evaluate

$$\mathrm{H}\left[p(x)\right] = 2 \int_0^{\sqrt{3}\sigma} \frac{1}{2\sqrt{3}\sigma} \ln \frac{1}{2\sqrt{3}\sigma}\, dx = \ln 2\sqrt{3}\sigma$$

.

Let's briefly compare the results of our computations as follows:

$$
\begin{aligned}
H_{\text{Gaussian}}(\sigma) &= 1/2 + \ln \sqrt{2\pi}\sigma = \frac{1}{2} + \ln \sqrt{2\pi} + \ln \sigma \approx 1.4189 + \ln \sigma \\
H_{\text{Triangle}}(\sigma) &= 1/2 + \ln \sqrt{6}\sigma = \frac{1}{2} + \ln \sqrt{6} + \ln \sigma \approx 1.3959 + \ln \sigma \\
H_{\text{Uniform}}(\sigma) &= \ln 2\sqrt{3}\sigma = 0 + \ln 2\sqrt{3} + \ln \sigma \approx 1.2425 + \ln \sigma
\end{aligned}
$$

This verifies that out of the three distributions, the Gaussian has the maximal entropy, as was expected.

Problem 2.23

a) To solve this problem, we need to find the inverse matrix, the determinant, and $\boldsymbol{w} = \boldsymbol{x} - \boldsymbol{\mu}$.

$$
\boldsymbol{\Sigma}^{-1} = \frac{1}{21}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & -2 \\ 0 & -2 & 5 \end{pmatrix} \qquad \det \boldsymbol{\Sigma} = 21 \qquad \boldsymbol{w} = \boldsymbol{x} - \boldsymbol{\mu} = \begin{pmatrix} -0.5 \\ -2 \\ -1 \end{pmatrix}
$$

The number of dimension d is 3. The solution is

$$
p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{3}{2}} 21^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{w}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{w}\right) = \frac{1}{(2\pi)^{\frac{3}{2}} 21^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\frac{1}{21}\frac{69}{4}\right).
$$

b) The eigenvectors of $\boldsymbol{\Sigma}$ are $\lambda_1 = 3$, $\lambda_1 = 7$ and $\lambda_1 = 21$. The corresponding eigenvectors are $\boldsymbol{v}_1 = (0, 1, -1)^T/\sqrt{2}$, $\boldsymbol{v}_2 = (0, 1, 1)^T/\sqrt{2}$ and $\boldsymbol{v}_3 = (1, 0, 0)^T$. The whitening transformation is

$$
\boldsymbol{A}_w = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2} = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 0 & \sqrt{2} \\ 1 & 1 & 0 \\ -1 & 1 & 0 \end{pmatrix}\begin{pmatrix} -\sqrt{3} & 0 & 0 \\ 0 & -\sqrt{7} & 0 \\ 0 & 0 & -\sqrt{21} \end{pmatrix}.
$$

The rest of the numerical computations are skipped.

c) Skipped.

d) Skipped.

e) We are going to examine if the p.d.f is unchanged when vectors are transformed with $\boldsymbol{T}^T\boldsymbol{x}$ and matrices with $\boldsymbol{T}^T\boldsymbol{\Sigma}\boldsymbol{T}$. Let's consider the term $(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ in the exponent first. Substituting $\boldsymbol{x} \mapsto \boldsymbol{T}^T\boldsymbol{x}$, $\boldsymbol{\mu} \mapsto \boldsymbol{T}^T\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} \mapsto \boldsymbol{T}^T\boldsymbol{\Sigma}\boldsymbol{T}$, we see that

$$
\begin{gathered}
\left(\boldsymbol{T}^T\boldsymbol{x} - \boldsymbol{T}^T\boldsymbol{\mu}\right)^T \left(\boldsymbol{T}^T\boldsymbol{\Sigma}\boldsymbol{T}\right)^{-1} \left(\boldsymbol{T}^T\boldsymbol{x} - \boldsymbol{T}^T\boldsymbol{\mu}\right) \\
\left(\boldsymbol{T}^T(\boldsymbol{x} - \boldsymbol{\mu})\right)^T \left(\boldsymbol{T}^T\boldsymbol{\Sigma}\boldsymbol{T}\right)^{-1} \boldsymbol{T}^T(\boldsymbol{x} - \boldsymbol{\mu}) \\
(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{T} \left(\boldsymbol{T}^T\boldsymbol{\Sigma}\boldsymbol{T}\right)^{-1} \boldsymbol{T}^T(\boldsymbol{x} - \boldsymbol{\mu}) \\
(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{T}\boldsymbol{T}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{T}^{-T}\boldsymbol{T}^T(\boldsymbol{x} - \boldsymbol{\mu}) \\
(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}),
\end{gathered}
$$

where we have used $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$ and $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$ from linear algebra. The density remains proportional when applying a linear transformation, but not unscaled, since the proportionality term $|\boldsymbol{\Sigma}|^{1/2}$ becomes $\left|\boldsymbol{T}^T\boldsymbol{\Sigma}\boldsymbol{T}\right|^{1/2} = \left|\boldsymbol{T}^T\right|^{1/2}|\boldsymbol{\Sigma}|^{1/2}|\boldsymbol{T}|^{1/2} = |\boldsymbol{T}|\,|\boldsymbol{\Sigma}|^{1/2}$.

f) Here we use the eigendecomposition of a symmetric matrix. We assume that $\boldsymbol{\Sigma}$ is positive definite such that every eigenvalue is positive. We write $\boldsymbol{\Sigma} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T$ and apply the whitening transformation.

$$\boldsymbol{A}_{\boldsymbol{w}}^T\boldsymbol{\Sigma}\boldsymbol{A}_{\boldsymbol{w}} = \boldsymbol{A}_{\boldsymbol{w}}^T\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T\boldsymbol{A}_{\boldsymbol{w}} = \left(\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}\right)^T\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T\left(\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}\right)$$

The matrix $\boldsymbol{\Phi}$ is orthogonal, so the transpose is the inverse. Using this fact and processing, we obtain

$$\left(\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}\right)^T\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T\left(\boldsymbol{\Phi}\boldsymbol{\Lambda}^{1/2}\right) = \left(\boldsymbol{\Lambda}^{-1/2}\right)^T\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1/2} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1/2} = \boldsymbol{I},$$

so the covariance is proportional to the identity matrix. The normalization constant becomes 1 , since the proportionality term becomes $|\boldsymbol{T}|\,|\boldsymbol{\Sigma}|^{1/2}$ under the transformation, and $|\boldsymbol{T}|\,|\boldsymbol{\Sigma}|^{1/2} = \left|\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}\right||\boldsymbol{\Sigma}|^{1/2} = \left|\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}\right|\left|\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T\right|^{1/2} = |\boldsymbol{I}| = 1$.

Problem 2.28

a) We prove that if $p(x_i-\mu_i, x_j-\mu_j) = p(x_i-\mu_i)p(x_j-\mu_j)$, then $\sigma_{ij} = \mathbb{E}\left[(x_i-\mu_i)(x_j-\mu_j)\right] = 0$. With words: we prove that statistical independence implies zero covariance.

$$\begin{gathered}\mathbb{E}\left[(x_i-\mu_i)(x_j-\mu_j)\right] = \\ \iint p(x_i-\mu_i, x_j-\mu_j)(x_i-\mu_i)(x_j-\mu_j)\,dx_jdx_i = \\ \iint p(x_i-\mu_i)p(x_j-\mu_j)(x_i-\mu_i)(x_j-\mu_j)\,dx_jdx_i \\ \int p(x_i-\mu_i)(x_i-\mu_i)\left(\int p(x_j-\mu_j)(x_j-\mu_j)\,dx_j\right)dx_i\end{gathered}$$

If the term in the parenthesis is identically zero, then $\sigma_{ij} = 0$. This is indeed true, since we find that

$$\int p(x_j-\mu_j)(x_j-\mu_j)\,dx_j = \mathbb{E}\left[(x_j-\mu_j)\right] = \mathbb{E}\left[x_j\right] - \mathbb{E}\left[\mu_j\right] = \mu_j - \mu_j = 0.$$

b) We wish to prove the converse of a) in the Gaussian case. To achieve this, we must show that $\sigma_{ij} = 0$ when $p(x_i-\mu_i, x_j-\mu_j) = p(x_i-\mu_i)p(x_j-\mu_j)$. Let's simplify the notation to x and y instead of x_i and x_j. If $\sigma_{xy} = 0$, then the covariance matrix is a diagonal matrix $\boldsymbol{D} = \operatorname{diag}(\sigma_x^2, \sigma_y^2)$. We write the probability $p(x_i-\mu_i, x_j-\mu_j)$ as $p(x,y)$, where the means μ_x and μ_y are both zero. We write

$$\begin{aligned}p(x,y) &= \frac{1}{(2\pi)^{2/2}\sigma_x\sigma_y}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{D}^{-1}\boldsymbol{x}\right) = \frac{1}{(2\pi)^{2/2}\sigma_x\sigma_y}\exp\left(-\frac{1}{2}\left(x^2/\sigma_x^2 + y^2/\sigma_y^2\right)\right) \\ &= \frac{1}{(2\pi)^{1/2}\sigma_x}\exp\left(-\frac{1}{2}\left(x^2/\sigma_x^2\right)\right)\cdot\frac{1}{(2\pi)^{1/2}\sigma_y}\exp\left(-\frac{1}{2}\left(y^2/\sigma_y^2\right)\right) = p(x)p(y).\end{aligned}$$

This proves that when $\sigma_{xy} = 0$, the covariance matrix is diagonal, and the Gaussian factors into products and we have statistical independence.

c) This problem asks us to find a counterexample of the above, i.e. an example showing that $\sigma_{xy} \not\Rightarrow p(x,y) = p(x)p(y)$. The probability density function

$$p(x,y) = K\frac{1}{1+x^2+y^2}, \quad K^{-1} = \iint_{\mathbb{R}} \frac{1}{1+x^2+y^2}\,dxdy$$

achieves this. The correlation is zero, since $\sigma_{xy} = \mathbb{E}\left[(x-0)(y-0)\right] = \iint_{\mathbb{R}} \frac{xy}{1+x^2+y^2}\,dxdy = \iint_{\mathbb{R}} I(x,y)\,dxdy$ is zero because the integrand $I(x,y)$ is an odd function. On the other hand, $p(x,y)$ does *not* factor into $p(x)p(y)$. We have proved that $\sigma_{xy} \not\Rightarrow p(x,y) = p(x)p(y)$ by finding a counterexample.

Problem 2.31

a) We'll assume that $\mu_1 < \mu_2$. Since $\sigma_1 = \sigma_2 = \sigma$, the minimum probability of error is achieved when setting $x^* = (\mu_1 + \mu_2)/2$. To follow the derivation below, it helps to draw the real line and two Gaussians. The probability of error is then

$$\begin{aligned}
P_e &= P(x \in R_2, \omega_1) + P(x \in R_1, \omega_2) \\
&= P(x \in R_2|\omega_1)P(\omega_1) + P(x \in R_1|\omega_2)P(\omega_2) \\
&= \int_{R_2} p(x|\omega_1)P(\omega_1)\,dx + \int_{R_1} p(x|\omega_2)P(\omega_2)\,dx \\
&= \frac{1}{2}\left(\int_{x^*}^{\infty} p(x|\omega_1)\,dx + \int_0^{x^*} p(x|\omega_2)\,dx\right) = \int_{x=(\mu_1+\mu_2)/2}^{\infty} p(x|\omega_1)\,dx \\
&= \int_{x=(\mu_1+\mu_2)/2}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2}\right)\,dx.
\end{aligned}$$

Changing variables to $u = (x-\mu_1)/\sigma$ and using $dx = \sigma\,du$ yields

$$P_e = \int_{u=a}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-u^2/\sigma^2\right)\,du,$$

where $a = (x-\mu_1)/\sigma = ((\mu_1+\mu_2)/2 - \mu_1)/\sigma = (\mu_2-\mu_1)/2\sigma$, as required.

b) Using the inequality stated in the problem, it remains to show that

$$\lim_{a\to\infty} f(a) = \lim_{a\to\infty} \frac{1}{\sqrt{2\pi}a} \exp\left(-a^2/\sigma^2\right) = 0.$$

This holds if the derivative is negative as $a \to \infty$, since then the function decreases as $a \to \infty$. The derivative of $f(a)$ is

$$f'(x) = -\exp\left(-a^2/2\right)\left(1 - \frac{1}{a^2}\right),$$

which is negative as long as $|a| \geq 1$. Alternatively, we see that both factors in $f(a)$ go to zero as $a \to \infty$.

Problem 2.43

a) p_{ij} is the probability that the i'th entry in the vector $\boldsymbol{x}$ equals 1, given a state of nature ω_j.
b) We decide ω_j if $P(\omega_j|\boldsymbol{x})$ is greater than $P(\omega_k|\boldsymbol{x})$ for every $k \neq j$.

$$P(\omega_j|\boldsymbol{x}) \propto p(\boldsymbol{x}|\omega_j)P(\omega_j)$$

We use the fact that $p(\boldsymbol{x}|\omega_j) = \prod_{i=1}^{d} p(x_i|\omega_j)$ since the entries are statistically independent. Furthermore, se wee that

$$p(x_i|\omega_j) = \begin{Bmatrix} p_{ij} & \text{if } x_i = 1 \\ 1 - p_{ij} & \text{if } x_i = 0 \end{Bmatrix} = p_{ij}^{x_i} \left(1 - p_{ij}\right)^{1-x_i}.$$

Now we take logarithms and obtain

$$\begin{aligned} \ln\left(\prod_{i=1}^{d} p(x_i|\omega_j)P(\omega_j)\right) &= \sum_{i=1}^{d} \ln p(x_i|\omega_j) + \ln P(\omega_j) \\ &= \sum_{i=1}^{d} \ln p_{ij}^{x_i} \left(1 - p_{ij}\right)^{1-x_i} + \ln P(\omega_j) \\ &= \sum_{i=1}^{d} x_i \ln p_{ij} + (1 - x_i)\ln(1 - p_{ij}) + \ln P(\omega_j), \end{aligned}$$

which is easily arranged to correspond with the expression in the problem statement. In summary we choose the class ω_j if the probability of that class given the data point exceeds the probability of every other data point.

2.2 Maximum-likelihood and Bayesian parameter estimation

Problem 3.2

a) The maximum likelihood estimate for θ is $\max_\theta p(x|\theta) = \max_\theta \prod_{i=1}^{n} p(x_i|\theta)$. The probability $p(x_i|\theta)$ is given by the expression

$$p(x_i|\theta) = \begin{cases} 1/\theta & \text{if } 0 \leq x_i \leq \theta \\ 0 & \text{if } x_i > \theta \end{cases}$$

The entire product $\prod_{i=1}^{n} p(x_i|\theta)$ is zero if any x_i is larger than θ, since then the corresponding factor is zero. Thus θ must be larger than, or equal to, $\max_k x_k$.

On the other hand, the product equals $\frac{1}{\theta^n}$, and taking logarithms we obtain $-k \ln \theta$. This function is maximized when θ is as small as possible.

The conclusion is that θ (1) must be $\geq \max_k x_k$ to avoid the likelihood being zero, and (2) as small as possible to maximize the likelihood. Therefore we choose $\hat{\theta} = \max_k x_k = \max \mathcal{D}$.
b) Skipping this plot. The explanation of why the other points are not needed is given in part a).

Problem 3.4

The maximum likelihood estimate is $p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{k=1}^{n} \prod_{i=1}^{d} \theta_i^{x_{ik}} (1-\theta_i)^{(1-x_{ik})}$. The log likelihood $\ell(\boldsymbol{\theta})$ is $\ln p(\mathcal{D}|\boldsymbol{\theta})$, which becomes

$$\ell(\boldsymbol{\theta}) = \sum_{k=1}^{n} \sum_{i=1}^{d} x_{ik} \ln \theta_i + (1 - x_{ik}) \ln (1 - \theta_i) \, .$$

Differentiating $\ell(\boldsymbol{\theta})$ with respect to component θ_i, every term in the $\sum_{i=1}^{d}$ vanishes except the i'th. We perform the differentiation and equate the result to zero, yielding

$$\frac{d\ell(\boldsymbol{\theta})}{\theta_i} = \sum_{k=1}^{n} \left[\frac{x_{ik}}{\theta_i} + \frac{x_{ik} - 1}{1 - \theta_i} \right] = \sum_{k=1}^{n} [x_{ik} - \theta_i] = 0.$$

Solving this for θ_i yields $\theta_i = \frac{1}{n} \sum_{k=1}^{n} x_{ik}$, or in vector notation, $\boldsymbol{\theta} = \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{x}_k$.

Problem 3.6

a) sdf

Problem 3.9

a) sdf

Problem 3.15

a) sdf

Problem 3.18

a) sdf

Problem 3.23

a) sdf

Problem 3.30

a) sdf

Problem 3.31

a) sdf

Problem 3.34

a) sdf

Problem 3.41

a) sdf

Problem 3.48

a) sdf

TEST

a) sdf