Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 1

Exercise 1: Estimating the Bayes Error $(10 + 10 + 10$ P)

The Bayes decision rule for the two classes classification problem results in the Bayes error

$$P(\text{error}) = \int P(\text{error}|\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x},$$

where $P(\text{error}|\boldsymbol{x}) = \min\left[\, P(\omega_1|\boldsymbol{x})\,,\ P(\omega_2|\boldsymbol{x})\,\right]$ is the probability of error for a particular input $\boldsymbol{x}$. Interestingly, while class posteriors $P(\omega_1|\boldsymbol{x})$ and $P(\omega_2|\boldsymbol{x})$ can often be expressed analytically and are integrable, the error function has discontinuities that prevent its analytical integration, and therefore, direct computation of the Bayes error.

(a) *Show* that the full error can be upper-bounded as follows:

$$P(\text{error}) \leq \int \frac{2}{\frac{1}{P(\omega_1|\boldsymbol{x})} + \frac{1}{P(\omega_2|\boldsymbol{x})}}\, p(\boldsymbol{x})\, d\boldsymbol{x}.$$

Note that the integrand is now continuous and corresponds to the harmonic mean of class posteriors weighted by $p(\boldsymbol{x})$.

(b) *Show* using this result that for the univariate probability distributions

$$p(x|\omega_1) = \frac{\pi^{-1}}{1 + (x-\mu)^2} \quad \text{and} \quad p(x|\omega_2) = \frac{\pi^{-1}}{1 + (x+\mu)^2},$$

the Bayes error can be upper-bounded by:

$$P(\text{error}) \leq \frac{2\, P(\omega_1)P(\omega_2)}{\sqrt{1 + 4\mu^2 P(\omega_1)P(\omega_2)}}$$

(Hint: you can use the identity $\int \frac{1}{ax^2+bx+c}\, dx = \frac{2\pi}{\sqrt{4ac-b^2}}$ for $b^2 < 4ac$.)

(c) *Explain* how you would estimate the error if there was no upper-bounds that are both tight and analytically integrable. Discuss following two cases: (1) the data is low-dimensional and (2) the data is high-dimensional.

Exercise 2: Bayes Decision Boundaries $(15 + 15$ P)

One might speculate that, in some cases, the generated data $p(x|\omega_1)$ and $p(x|\omega_2)$ is of no use to improve the accuracy of a classifier, in which case one should only rely on prior class probabilities $P(\omega_1)$ and $P(\omega_2)$ assumed here to be strictly positive.

For the first part of this exercise, we assume that the data for each class is generated by the univariate Laplacian probability distributions:

$$p(x|\omega_1) = \frac{1}{2\sigma} \exp\Big(-\frac{|x-\mu|}{\sigma}\Big) \quad \text{and} \quad p(x|\omega_2) = \frac{1}{2\sigma} \exp\Big(-\frac{|x+\mu|}{\sigma}\Big).$$

where $\mu, \sigma > 0$.

(a) *Determine* for which values of $P(\omega_1), P(\omega_2), \mu, \sigma$ the optimal decision is to always predict the first class (i.e. under which conditions $P(\text{error}|x) = P(\omega_2|x)\ \ \forall\, x \in \mathbb{R}$).

(b) *Repeat* the exercise for the case where the data for each class is generated by the univariate Gaussian probability distributions:

$$p(x|\omega_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big) \quad \text{and} \quad p(x|\omega_2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x+\mu)^2}{2\sigma^2}\Big).$$

where $\mu, \sigma > 0$.

Exercise 3: Programming (40 P)

Download the programming files on ISIS and follow the instructions.

Machine Learning 1

Week 1

$\boxed{1}$ a) $P(\text{error}) = \int P(\text{error} \mid \vec{x}) \cdot p(\vec{x}) \cdot d\vec{x}$

$$= \int \min\left[P(\omega_1 \mid \vec{x}), P(\omega_2 \mid \vec{x})\right] \; p(\vec{x}) \cdot d\vec{x}$$

1) $$= \int M_{-\infty}\left(P(\omega_1 \mid \vec{x}), P(\omega_2 \mid \vec{x})\right) p(\vec{x})\, d\vec{x}$$

2) $$\leq \int M_{-1}\left(P(\omega_1 \mid \vec{x}), P(\omega_2 \mid \vec{x})\right) p(\vec{x})\, d\vec{x}$$

3) $$= \int \frac{2}{\frac{1}{P(\omega_1 \mid \vec{x})} + \frac{1}{P(\omega_2 \mid \vec{x})}} \, p(\vec{x})\, d\vec{x}$$

with 1) $M_p(x_1, \dots, x_n) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^{\,p}\right)^{\frac{1}{p}}$

2) $M_p \leq M_q$ if $p < q$

3) $M_{-1}\left(P(\omega_1 \mid \vec{x}), P(\omega_2 \mid \vec{x})\right) = \left(\frac{1}{2} P(\omega_1 \mid \vec{x})^{-1} + \frac{1}{2} P(\omega_2 \mid x)^{-1}\right)^{-1}$

$$= \frac{1}{\frac{1}{2}\frac{1}{P(\omega_1 \mid \vec{x})} + \frac{1}{2}\frac{1}{P(\omega_2 \mid \vec{x})}} = \frac{2}{\frac{1}{P(\omega_1 \mid \vec{x})} + \frac{1}{P(\omega_2 \mid \vec{x})}}$$

b)

$$P(\text{error}) \leq \int \frac{2}{\frac{1}{P(\omega_1|\vec{x})} + \frac{1}{P(\omega_2|\vec{x})}} \, p(\vec{x}) \, d\vec{x}$$

Bayes rule:

$$P(\omega_j|\vec{x}) = \frac{p(\vec{x}|\omega_j)P(\omega_j)}{p(\vec{x})}$$

$$= \int \frac{2}{\frac{1}{p(\vec{x}|\omega_1)P_1} + \frac{1}{p(\vec{x}|\omega_2)P_2}} \, d\vec{x}$$

$$P(\omega_j) = P_j$$

$$p(\vec{x}|\omega_1) = \frac{\pi^{-1}}{1+(x-\mu)^2}$$

$$p(\vec{x}|\omega_2) = \frac{\pi^{-1}}{1+(x+\mu)^2}$$

$$= \int \frac{2}{\frac{1}{\frac{\pi^{-1}P_1}{1+(x-\mu)^2}} + \frac{1}{\frac{\pi^{-1}P_2}{1+(x+\mu)^2}}} \, d\vec{x}$$

$$= \int \frac{2P_1P_2\pi^{-1}}{P_2(1+(x-\mu)^2) + P_1(1+(x+\mu)^2)}$$

$$= 2P_1P_2\pi^{-1} \int \frac{1}{\underbrace{1\vec{x}^2}_{a} + \underbrace{2\mu(P_1-P_2)}_{b}\vec{x} + \underbrace{1+\mu^2}_{c}} \, d\vec{x}$$

$$\int \frac{1}{ax^2+bx+c} \, dx = \frac{2\pi}{\sqrt{4ac-b^2}}$$

In our case

$$(2\mu(P_1-P_2))^2 < 4(1+\mu^2)$$

because $(P_1-P_2)^2 < 1$

$$= 2P_1P_2 \frac{2\pi}{\sqrt{4+4\mu^2-4\mu^2(P_1-P_2)^2}}$$

$$= 2P_1P_2 \frac{1}{\sqrt{4+16\mu^2P_1P_2}}$$

$$= \frac{2P_1P_2}{\sqrt{1+4\mu^2P_1P_2}}$$

c) 1) integrate the original error function $P(\text{error}|\vec{x})$ numerically for low dimensional data

2) Create discriminants $g_1(x), g_2(x)$ supporting an optimal classifier and sample hierarchically $\omega_j \sim [P(\omega_1), P(\omega_2)]$ and $P(\vec{x}|\omega_j)$ and compute the average.

$\boxed{2}$ a) For Bayes optimal decision boundary we need two discriminants

$$g_1(x) = -\frac{|x-\mu|}{\sigma} + \log P(w_1)$$

$$g_2(x) = -\frac{|x+\mu|}{\sigma} + \log P(w_2)$$

Next, we need to check for which parameter values and all x $\forall_x : g_1(x) \geq g_2(x)$ holds. We have

$$x \leq -\mu \Rightarrow 2\mu \leq \sigma \cdot [\log P(w_1) - \log P(w_2)]$$

$$-\mu < x < \mu \Rightarrow 2x \leq \sigma [\log P(w_1) - \log P(w_2)]$$

$$x \geq \mu \Rightarrow -2\mu \leq \sigma [\log P(w_1) - \log P(w_2)]$$

From this set of equations we can infer the parameter for w_1 prediction:

$$\left\{ (P(w_1), P(w_2), \mu, \sigma) \,\middle|\, \log \frac{P(w_1)}{P(w_2)} \geq \frac{2\mu}{\sigma} \right\}$$

b) The new pair of discriminants is

$$g_1(x) = -\frac{(x-\mu)^2}{2\sigma^2} + \log P(w_1)$$

$$g_2(x) = -\frac{(x+\mu)^2}{2\sigma^2} + \log P(w_2)$$

The inequality $g_1(x) \geq g_2(x)$ becomes

$$-2x\mu \leq \sigma [\log P(w_1) - \log P(w_2)]$$

There is no set of parameters that always predicts class w_1.