STATISTICAL MACHINE LEARNING (WS2025/26)
SEMINAR: PREDICTOR EVALUATION

Assignment 1. Consider a binary classification problem with scalar observation $\mathcal{X} = \mathbb{R}$, two possible classes $\mathcal{Y} = \{-1, +1\}$ and the 0/1-loss $\ell(y, y') = [\![y \neq y']\!]$. The observations for both classes are generated according to Gaussian distributions. Specifically, the joint probability distribution of the observation $x \in \mathbb{R}$ and tha class label $y \in \mathcal{Y}$ is given by:

$$p(x, y) = p(y)\,\frac{1}{\sqrt{2\pi}\sigma_y}\exp\Big(-\frac{1}{2\sigma_y^2}(x - \mu_y)^2\Big)\,, \qquad y \in \mathcal{Y}\,,$$

where $p(y)$ is the prior distribution of the class y, μ_+ and μ_- are the means of the distributions for $y = +1$ and $y = -1$, respectively, $\sigma_+ > 0$ and $\sigma_- > 0$ are the corresponding standard deviations.

a) Assume $\mu_- < \mu_+$ and $\sigma_+ = \sigma_-$. Show that, under these conditions, the Bayes optimal prediction strategy is a thresholding rule of the form:

$$h(x) = \begin{cases} -1 & \text{if} \quad x < \theta\,, \\ +1 & \text{if} \quad x \geq \theta\,, \end{cases}$$

where $\theta \in \mathbb{R}$ is a scalar theresold. Derive explicit formula for computing θ.

b) Now assume $\mu_+ = \mu_-$ and $\sigma_+ \neq \sigma_-$. Determine the optimal prediction strategy under these conditions.

c*) Finally, consider the case where $\mu_+ = \mu_-$, $\sigma_+ \neq \sigma_-$, and both classes have nonzero prior probabilities, i.e. $p(+1) > 0$ and $p(-1) > 0$. Is it possible for the Bayes classifier to assign all inputs $x \in \mathbb{R}$ to a single class? Prove your answer.

Solution 1. The Bayes classifier in case of the 0/1-loss and two classes assigns the input x into the class with the higher class posterior $p(y \mid x)$, or equivalently with the higher $p(x, y)$, that is,

$$h^*(x) = \begin{cases} +1 & \text{if} \quad p(x, y = +1) > p(x, y = -1) \\ -1 & \text{if} \quad p(x, y = +1) < p(x, y = -1) \end{cases}$$

Note that the boundary inputs, $p(x, y = +1) = p(x, y = -1)$, can be assigned to an arbitrary class. Let us define a discriminant function $f(x)$ as a logarithm of the likelihood ratio:

$$f(x) = \log\left(\frac{p(x, y = +1)}{p(x, y = -1)}\right)\,. \tag{1}$$

The Bayes classifier can be expressed equaivalently as the sign of the discriminant function:

$$h^*(x) = \operatorname{sign}(f(x))\,.$$

After substituting

$$p(x, y) = p(y)\,\frac{1}{\sqrt{2\pi}\sigma_y}\exp\Big(-\frac{1}{2\sigma_y^2}(x - \mu_y)^2\Big)$$

to (1) we get

$$
\begin{aligned}
f(x) &= \log p(x, y=+1) - \log p(x, y=-1) \\
&= \log \frac{p(+1)}{\sigma_+} - \log \sqrt{2\pi} - \frac{(x-\mu_+)^2}{2\sigma_+^2} - \log \frac{p(-1)}{\sigma_-} + \log \sqrt{2\pi} + \frac{(x-\mu_-)^2}{2\sigma_-^2} \\
&= x^2 \underbrace{\left(\frac{1}{2\sigma_-^2} - \frac{1}{2\sigma_+^2} \right)}_{a} + x \underbrace{\left(\frac{\mu_+}{\sigma_+^2} - \frac{\mu_-}{\sigma_-^2} \right)}_{b} + \underbrace{\frac{\mu_-^2}{2\sigma_-^2} - \frac{\mu_+^2}{2\sigma_+^2} + \log \frac{p(+1)\sigma_-}{p(-1)\sigma_+}}_{c} \\
&= a\, x^2 + b\, x + c\,.
\end{aligned}
$$

a) In case of $\sigma_+ = \sigma_- = \sigma$, the multiplier in fron of x^2 is $a = 0$ and the discriminant function becomes a linear function. Hence,

$$
h^*(x) = \operatorname{sign}(f(x)) = \operatorname{sign}(b\, x + c) = \begin{cases} -1 & \text{if} \quad x < \theta \\ +1 & \text{if} \quad x \geq \theta \end{cases}
$$

where θ is a solution of the linear equation $f(x) = b\, x + c = 0$, i.e.

$$
\theta = -\frac{c}{b} = \frac{\mu_+ + \mu_-}{2} + \frac{\sigma^2}{\mu_- - \mu_+} \log \frac{p(+1)}{p(-1)}\,.
$$

b) In case of $\mu_+ = \mu_- = \mu$ we can rewrite the discriminant function as

$$
f(x) = \frac{1}{2} \left(\frac{1}{\sigma_-^2} - \frac{1}{\sigma_+^2} \right) (x - \mu)^2 + \log \frac{p(+1)\sigma_-}{p(-1)\sigma_+}\,.
$$

If $\sigma_- < \sigma_+$, the discriminant function is convex. If $p(+1)\sigma_- < p(-1)\sigma_+$, then the minimum of $f(x)$ is a negative number and the quadratic equation $f(x) = 0$ has two solutions, which we denote θ_1 and θ_2. In this case, the Bayes classifier assigns the inputs x that fall to the interval $[\theta_1, \theta_2]$ into the negative class and the inputs outside the interval to the positive class. If $p(+1)\sigma_- < p(-1)\sigma_+$, the minimum of $f(x)$ is a positive number and the Bayes classifier assigns all inputs to the negative class. If $\sigma_- > \sigma_+$, the discriminant function is concave and the analysis is analogous.

c*) Yes, it can happen. For example, when $\mu_+ = \mu_-$, $\sigma_- < \sigma_+$ and $p(+1)\sigma_- > p(-1)\sigma_+$, the discriminant function attains a positive values for all x and hence all inputs are assigned to the positive class.

Assignment 2. We are given a prediction strategy $h\colon \mathcal{X} \to \mathcal{Y} = \{1, \ldots, Y\}$ assigning observations $x \in \mathcal{X}$ into one of Y classes. Our task is to estimate the true error $R(p, h) = \mathbb{E}_{(x,y)\sim p} \ell(y, h(x))$ where $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a chosen loss function. To this end, we collect a test set $S_n = ((x_i, y_i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, n)$ i.i.d. drawn from the distribution $p(x, y)$, compute the test error $\hat{R}(S_n, h) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, h(x_i))$ and use it to construct the confidence interval such that

$$
R(p, h) \in \left(\hat{R}(S_n, h) - \varepsilon, \hat{R}(S_n, h) + \varepsilon \right) \quad \text{holds with probability } 1-\delta \in (0, 1) \text{ at least.} \tag{2}
$$

The number of test examples $n \in \mathbb{N}$, the error margin $\varepsilon > 0$ and the confidence level $1 - \delta \in (0, 1)$ are three interdependent variables, i.e., fixing two of the variables allows to compute the third one.

a) Use the Hoeffding's inequality to derive a formula to compute ε as a function of n and δ such that (2) holds.

b) Use the Hoeffding's inequality to derive a formula to compute n as a function of ε and δ such that (2) holds.

c) Instantiate the formulas derived in a) and b) for the following loss functions:

(1) $\ell(y, y') = [\![y \neq y']\!]$
(2) $\ell(y, y') = |y - y'|$
(3) $\ell(y, y') = [\![|y - y'| \geq K]\!]$ where $K < Y$.

d) Assume that we use the loss $\ell(y, y') = [\![y \neq y']\!]$. Plot the error margin ε as a function of the number of examples $n \in \{10, 100, \ldots, 100000\}$ for $\delta \in \{0.1, 0.05, 0.01\}$.

e) Assume that we use the 0/1-loss $\ell(y, y') = [\![y \neq y']\!]$. What is the minimal number of examples n we need to use to have a guarantee that the test error will approximate the generalization error $\pm 1\%$ with probability 95% at least?

Solution 2. **a)** Formula for $\varepsilon(n, \delta)$. Assume the loss values are bounded in an interval $[\ell_{\min}, \ell_{\max}]$. Let $Z_i = \ell(y_i, h(x_i)) \in [\ell_{\min} - \ell_{\max}]$, $\hat{R} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ and $R = \mathbb{E}_{S_n \sim p^n}[\ell(y, h(x)]$. Hoeffding's inequality says

$$\mathbb{P}\big(|\hat{R} - R| \geq \varepsilon\big) \leq 2\exp\Big(-\frac{2n\varepsilon^2}{(\ell_{\max} - \ell_{\min})^2}\Big).$$

We want $\mathbb{P}(|\hat{R} - R| < \varepsilon) \geq 1 - \delta$ which is the same as $\mathbb{P}(|\hat{R} - R| \geq \varepsilon) \leq \delta$. So, equivalently, we want the right-hand side of the Hoeffding inequality to be $\leq \delta$. Solve for ε:

$$2\exp\Big(-\frac{2n\varepsilon^2}{(\ell_{\max} - \ell_{\min})^2}\Big) \leq \delta \quad \iff \quad -\frac{2n\varepsilon^2}{(\ell_{\max} - \ell_{\min})^2} \leq \ln\frac{\delta}{2}$$

which gives

$$\boxed{\varepsilon(n, \delta) = (\ell_{\max} - \ell_{\min})\sqrt{\frac{1}{2n}\ln\big(\tfrac{2}{\delta}\big)}}.$$

With this ε we have $\mathbb{P}\big(|\hat{R} - R| \leq \varepsilon\big) \geq 1 - \delta$.

b) Formula for $n(\varepsilon, \delta)$. Rearrange the same bound to solve for n:

$$2\exp\Big(-\frac{2n\varepsilon^2}{(\ell_{\max} - \ell_{\min})^2}\Big) \leq \delta \quad \Longrightarrow \quad n \geq \frac{(\ell_{\max} - \ell_{\min})^2}{2\varepsilon^2}\ln\big(\tfrac{2}{\delta}\big).$$

So one can choose

$$\boxed{n(\varepsilon, \delta) = \frac{(\ell_{\max} - \ell_{\min})^2}{2\varepsilon^2}\ln\big(\tfrac{2}{\delta}\big)}$$

(or the ceiling of the right-hand side to get an integer).

c) Instantiate for the three losses:

(1) $\ell(y, y') = [\![y \neq y']\!]$. This loss takes values in $\{0, 1\}$, so $\ell_{\min} = 0, \ell_{\max} = 1$ and $\ell_{\max} - \ell_{\min} = 1$. Thus

$$\varepsilon = \sqrt{\frac{1}{2n}\ln\left(\tfrac{2}{\delta}\right)}, \qquad n = \frac{1}{2\varepsilon^2}\ln\left(\tfrac{2}{\delta}\right).$$

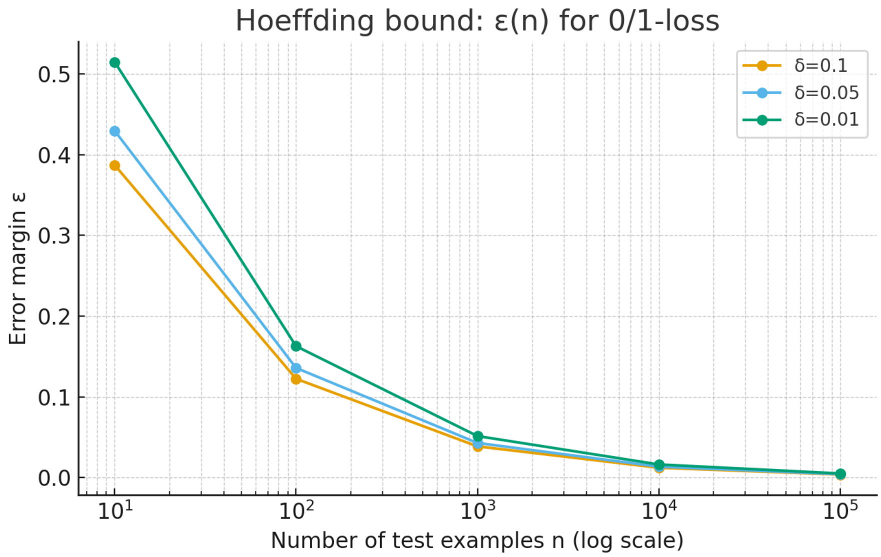
(2) $\ell(y, y') = |y - y'|$ with $y, y' \in \{1, \dots, Y\}$. The range is $[0, Y-1]$, so $\ell_{\max} - \ell_{\min} = Y - 1$. Hence

$$\varepsilon = (Y-1)\sqrt{\frac{1}{2n}\ln\left(\tfrac{2}{\delta}\right)}, \qquad n = \frac{(Y-1)^2}{2\varepsilon^2}\ln\left(\tfrac{2}{\delta}\right).$$

(3) $\ell(y, y') = [\![|y - y'| \geq K]\!]$. This is again a $\{0, 1\}$-valued loss, so the formulas are the same as in 1):

$$\varepsilon = \sqrt{\frac{1}{2n}\ln\left(\tfrac{2}{\delta}\right)}, \qquad n = \frac{1}{2\varepsilon^2}\ln\left(\tfrac{2}{\delta}\right).$$

d) Plot ε vs n for 0/1-loss and $\delta \in \{0.1, 0.05, 0.01\}$.



The curve shows the $O(1/\sqrt{n})$ decay and the dependence on δ via $O(\sqrt{\ln(2/\delta)})$.

e) Minimal n for 0/1-loss, $\varepsilon = 1\%$, $1 - \delta = 95\%$. We need $\varepsilon = 0.01$ and $\delta = 0.05$. Use $n \geq \frac{1}{2\varepsilon^2}\ln\left(\frac{2}{\delta}\right)$. Numerically:

$$n \geq \frac{1}{2 \cdot 0.01^2}\ln\left(\tfrac{2}{0.05}\right) = \frac{1}{2 \cdot 10^{-4}}\ln(40) = 5000 \cdot \ln(40) \approx 18444.3973.$$

Rounding up to an integer,

$$\boxed{n_{\min} = 18\,445 \text{ examples (approximately).}}$$

Assignment 3. Let $S_n = ((x_i, y_i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, n)$ be a test set i.i.d drawn from some $p(x, y)$ and let $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function. The test error $\hat{R}(S_n, h) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h(x_i))$ is an unbiased estimator of the true error $R(p, h) = \mathbb{E}_{(x,y)\sim p}\ell(y, h(x))$.

a) What does it mean that the test error is an unbiased estimator of the true error?

b) Prove that it holds true.

c) Prove that the variance of the test error decreases as $1/n$, i.e. more test examples reduces the estimator error.

Solution 3. a) a) Saying the test error $\hat{R}(S_n, h)$ is an unbiased estimator of the true error $R(p, h)$ means

$$\mathbb{E}_{S_n \sim p^n}\big[\hat{R}(S_n, h)\big] = R(p, h),$$

where the expectation is over the random draw of the test sample S_n. In words: on average (over repeated draws of test sets of size n) the test error equals the true risk — there is no systematic over- or under-estimation.

b) Let

$$z_i := \ell\big(y_i, h(x_i)\big), \qquad i = 1, \dots, n,$$

so $\hat{R}(S_n, h) = \dfrac{1}{n}\sum_{i=1}^n z_i$. The (x_i, y_i) are i.i.d., hence the z_i are i.i.d. with

$$\mathbb{E}[z_i] = \mathbb{E}_{(x,y)\sim p}\big[\ell(y, h(x))\big] = R(p, h).$$

Using linearity of expectation,

$$\mathbb{E}_{S_n \sim p^n}\big[\hat{R}(S_n, h)\big] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n Z_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i] = \frac{1}{n}\cdot n\, R(p, h) = R(p, h).$$

Thus $\hat{R}$ is unbiased.

(c) Again write $z_i = \ell(y_i, h(x_i))$, and set $\mu = \mathbb{E}_{S_n \sim p^n}[\hat{R}(S_n, h)]$. Because the z_i are independent, the variance reads

$$\begin{aligned}
\mathbb{V}_{S_n \sim p^n}\big[\hat{R}(S_n, h)\big] &= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n z_i\right] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n z_i - \mu\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n (z_i - \mu)(z_j - \mu)\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n (z_i - \mu)^2\right] + \frac{1}{n^2}\sum_{i \neq j} \underbrace{\mathbb{E}\left[(z_i - \mu)(z_j - \mu)\right]}_{=0 \text{ because } z_i \text{ and } z_j \text{ are independet}} \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}(z_i) \\
&= \frac{1}{n^2}\cdot n\, \mathbb{V}_{(x,y)\sim p}\big[\ell(y, h(x))\big].
\end{aligned}$$

So

$$\mathbb{V}\big[\hat{R}\big] = \frac{\sigma^2}{n} \qquad \text{where } \sigma^2 = \mathbb{V}_{(x,y)\sim p}\big[\ell(y, h(x))\big].$$

As a consequence, the variance decreases as $1/n$ — more test examples reduce estimator variance.