# MACHINE LEARNING FUNDAMENTALS - LS2026
# SEMINAR: EMPIRICAL RISK MINIMIZATION

CZECH TECHNICAL UNIVERSITY IN PRAGUE
V. FRANC

**Assignment 1.** Assume you want to learn a predictor $h\colon \mathcal{X} \to \mathcal{Y} = \{-1, +1\}$ that minimizes the probability of misclassification. You are given a training sample $T_m = \big((x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, m\big)$ drawn i.i.d. from an unknown distribution $p(x, y)$. A learning algorithm $A\colon \bigcup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ is employed, which returns a classifier $h_m = A(T_m)$ from the hypothesis class $\mathcal{H} = \big\{h(x) = \operatorname{sign}(\langle w, x\rangle + b) \mid w \in \mathbb{R}^n, b \in \mathbb{R}\big\}$. Determine the approximation error of the learning algorithm A in each of the following cases.

**a)** The input is a vector of binary features $x = (x_1, \ldots, x_n) \in \{0, 1\}^n$, and the data $(x, y)$ are generated according to the joint distribution

$$p(x, y) = p(y) \prod_{i=1}^{n} p(x_i \mid y),$$

where $p(x_i \mid y)$ are conditional probabilities of the binary features given the class label $y$, and $p(y)$ is the class prior.

**b)** The input is a real-valued vector $x \in \mathcal{X} = \mathbb{R}^n$, and the data $(x, y)$ are generated from the Gaussian mixture model

$$p(x, y) = p(y) \frac{1}{(2\pi)^{\frac{n}{2}} \det(C_y)^{\frac{1}{2}}} \exp\Big( -\tfrac{1}{2}(x - \mu_y)^\top C_y^{-1} (x - \mu_y)\Big),$$

where $\mu_y \in \mathbb{R}^n$ are the class means, $C_y \in \mathbb{R}^{n \times n}$ are the class covariance matrices, and $p(y)$ are the class priors. Assume $C_+ = C_-$.

**c)** Consider the setup from part b). Discuss whether the approximation error will increase or decrease if $C_+ \neq C_-$.

**Solution 1. a)** The approximation error is the difference between the best risk in the class $\mathcal{H}$ and the risk of the Bayes classifier $h_*$, i.e. the best attainable classifier. The Bayes classifier reads:

$$h_*(x) = \operatorname{sign}(f(x))$$

where

$$\begin{aligned} f(x) &= \log \frac{p(y=+1 \mid x)}{p(y=-1 \mid x)} = \log \frac{p(x, y=+1)}{p(x, y=-1)} \\ &= \log \frac{p(y=+1)}{p(y=-1)} + \sum_{i=1}^{n} \log \frac{p(x_i \mid y=+1)}{p(x_i \mid y=-1)} \\ &= \log \frac{p(y=+1)}{p(y=-1)} + \sum_{i=1}^{n} (1-x_i) \log \frac{p(x_i=0 \mid y=+1)}{p(x_i=0 \mid y=-1)} + \sum_{i=1}^{n} x_i \log \frac{p(x_i=1 \mid y=+1)}{p(x_i=1 \mid y=-1)} \\ &= b + \sum_{i=1}^{n} x_i \, w_i \, . \end{aligned}$$

Hence, the Bayes classifier is an instance of the linear classifier. This implies that the approximation error equals $0$.

**b )** Gaussian class-conditionals with equal covariances. Given

$$p(x \mid y) = \frac{1}{(2\pi)^{n/2} \det(C_y)^{1/2}} \exp\Big( -\tfrac{1}{2}(x-\mu_y)^\top C_y^{-1}(x-\mu_y)\Big)$$

and assume $C_+ = C_- = C$. The log-likelihood ratio is

$$\log \frac{p(y=+1 \mid x)}{p(y=-1 \mid x)} = \log \frac{p(y=+1)}{p(y=-1)} - \frac{1}{2}\Big[(x-\mu_+)^\top C^{-1}(x-\mu_+) - (x-\mu_-)^\top C^{-1}(x-\mu_-)\Big] + \text{const.}$$

Expand the quadratic forms; the $x^\top C^{-1} x$ terms cancel, leaving a linear function in $x$:

$$\text{const} + x^\top C^{-1}(\mu_+ - \mu_-).$$

Thus the Bayes decision is $h_*(x) = \operatorname{sign}(\langle w, x\rangle + b)$ with $w = C^{-1}(\mu_+ - \mu_-)$. So again the Bayes classifier is linear. That is, the Bayes classifier lies in $\mathcal{H}$, and therefor the appoximation error is zero in case $C_+ = C_-$.

**c)** What if $C_+ \neq C_-$ ? Write the log posterior odds:

$$\begin{aligned} \log \frac{p(y=1+ \mid x)}{p(y=-1 \mid x)} = \log \frac{p(y=+1)}{p(y=-1)} - \frac{1}{2}\big[&(x-\mu_+)^\top C_+^{-1}(x-\mu_+) - \\ &(x-\mu_-)^\top C_-^{-1}(x-\mu_-)\big] - \frac{1}{2}\log \frac{\det C_+}{\det C_-} + \text{const.} \end{aligned}$$

When $C_+^{-1} \neq C_-^{-1}$ the expression contains a genuine quadratic term in $x$:

$$-\frac{1}{2} x^\top (C_+^{-1} - C_-^{-1}) x + x^\top \big(C_+^{-1}\mu_+ - C_-^{-1}\mu_-\big) + \text{const.}$$

So the Bayes decision boundary is typically quadratic (a general conic). A general quadratic boundary cannot in general be represented by a linear classifier. Therefore the approximation error becomes positive.

**Assignment 2.** You aim to train a Convolutional Neural Network (CNN) classifier, $h\colon \mathcal{X} \to \mathcal{Y}$, to predict a digit $y \in \mathcal{Y} = \{0, 1, \ldots, 9\}$ from an image $x \in \mathcal{X}$. The objective is to minimize the probability of misclassification $R(p,h) = \mathbb{E}_{(x,y)\sim p}\big([\![\, y \neq h(x) \,]\!]\big)$ . The CNN is trained using the Stochastic Gradient Descent (SGD) algorithm for $100$ epochs. After each epoch, the

model's weights are saved, resulting in a set of 100 CNN classifiers: $\mathcal{H} = \{h_i \colon \mathcal{X} \to \mathcal{Y} \mid i = 1, \ldots, 100\}$ . At the end, in the model selection stage, the best CNN classifier is selected by

$$\hat{h} \in \arg\min_{h \in \mathcal{H}} \hat{R}(V_m, h) \,,$$

where the validation error is defined as:

$$\hat{R}(V_m, h) = \frac{1}{m} \sum_{i=1}^{m} [\![ y_i \neq h(x_i) ]\!] \,.$$

The validation set $V_m = ((x_i, y_i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m)$ consists of $m$ i.i.d. samples drawn from $p(x, y)$. The validation set $V_m$ is not used during training the predictors in $\mathcal{H}$, but it is used to select the best-performing claassifier among them.

Since no additional test data is available, we misclassification error $R(p, \hat{h})$ is estimated using the same validation error $\hat{R}(V_m, \hat{h})$.

**a)** Does the procedure used to select the best CNN model $\hat{h}$ constitute an application of the Empirical Risk Minimization (ERM) principle?

**b)** What is the minimum size of the validation set $V_m$ required to assert with $99\%$ confidence that the true error $R(p, \hat{h})$ differs from the validation error $\hat{R}(V_m, \hat{h})$ by at most $1\%$ ?

**c)** Suppose the validation set contains $m = 20,000$ examples. Compute the smallest $\varepsilon$ such that, with $95\%$ confidence,

$$\left| \hat{R}(V_m, \hat{h}) - R(p, \hat{h}) \right| \leq \varepsilon \,.$$

**Solution 2. a)** Yes, the model selection step is an application of the ERM principle over finite hypothesis space $\mathcal{H}$. The validation error $\hat{R}(V_m, h)$ is the empirical risk on the validation set $V_m$.

**b)** For $0/1$-loss, Hoeffding's inequality combined with the union bound yields the following unifomr deviation bound:

$$\mathbb{P}\left( \max_{h \in \mathcal{H}} \left| \hat{R}(T_m, h) - R(p, h) \right| \geq \varepsilon \right) \leq 2|\mathcal{H}| e^{-2m\varepsilon^2}$$

valied for any $\varepsilon > 0$. Setting the RHS equal to $\delta$ and solving for $m$, we obtain:

$$m \geq \frac{1}{2\varepsilon^2} \ln\left( \frac{2|\mathcal{H}|}{\delta} \right) \,.$$

Here, $|\mathcal{H}| = 100$, $\varepsilon = 0.01$, $\delta = 0.01$ (since $99\%$ confidence). Thus,

$$m \geq \frac{1}{2(0.01)^2} \ln\left( \frac{2 \cdot 100}{0.01} \right) = \frac{1}{0.0002} \ln(20000) \approx 49517.44 \,.$$

Therefore, the minimum required size of the validation set is $m = 49,518$.

**c)** Using the the same bound, we now solve for $\varepsilon$:

$$\varepsilon = \sqrt{\frac{1}{2m} \ln\left( \frac{2|\mathcal{H}|}{\delta} \right)} \,.$$

With $|\mathcal{H}| = 100$, $m = 20000$,$\delta = 0.05$ (corresponding to $95\%$ confidence), we obtain

$$\varepsilon = \sqrt{\frac{1}{2 \cdot 20000} \ln\left(\frac{200}{0.05}\right)} \approx 0.01440 \, .$$

Thus, with probability at least $95\%$,

$$|\hat{R}(V_m, \hat{h}) - R(p, \hat{h})| \leq 0.0144 \, .$$