STATISTICAL MACHINE LEARNING (WS2018)
SEMINAR 2

Remark: Tools needed for solving Assignment 3 and 4 are discussed at the end of the lecture on the ERM I which will be finished on Tuesday 16, 2018.

Assignment 1. Consider the task of age estimation based on visual cues. Let us denote the visual features by $x \in \mathcal{X}$ and the unknown age by $y \in \mathbb{N}$. The statistical relation between the two random variables is known and given by their joint distribution $p(x, y)$.
a) Deduce the optimal inference rule for the loss function $\ell(y, y') = |y - y'|^2$.
b) Same for the loss function $\ell(y, y') = |y - y'|$.

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 expected risk $R^\ell(h) = \mathbb{E}_{(x,y)\sim p}\ell(y, h(x))$ where $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is some application specific loss function. To this end, we collect a set of examples $\mathcal{S}^l = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, l\}$ drawn i.i.d. from the distribution $p(x, y)$ and compute the test error

$$R_{\mathcal{S}^l}(h) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^j, h(x^j)) \ .$$

What is the minimal number of test examples l we need to collect in order to have a guarantee that the expected risk $R^\ell(h)$ is inside the interval $(R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon)$ with probability $\gamma \in (0, 1)$ for some predefined $\varepsilon > 0$?

a) Give a formula to compute l as a function of ε and γ for the $0/1$-loss $\ell(y, y') = [\![y \neq y']\!]$. Evaluate l for $\varepsilon = 0.01$ and $\gamma \in \{0.90, 0.95, 0.99\}$.
b) Solve the problem a) in case that the loss is the mean absolute error, $\ell(y, y') = |y - y'|$. Evaluate l for $\varepsilon = 1$, $Y = 100$ and $\gamma \in \{0.90, 0.95, 0.99\}$.
c) How do the formulas depend on the particular loss function?

Assignment 3. We are given a set $\mathcal{H} = \{h_i\colon \mathcal{X} \to \{1, \ldots, 100\} \mid i = 1, \ldots, 1000\}$ containing 1000 strategies each predicting the human age $y \in \{1, \ldots, 100\}$ from a facial image $x \in \mathcal{X}$. The quality of a single strategy is measured by the expected absolute deviation between the predicted age and the true age

$$R^{\text{MAE}}(h) = \mathbb{E}_{(x,y)\sim p}(|y - h(x)|) \ ,$$

where the expectation is computed w.r.t. an unknown distribution $p(x, y)$. The empirical estimate of $R^{\text{MAE}}(h)$ reads

$$R_{\mathcal{T}^m}(h) = \frac{1}{m}\sum_{i=1}^{m} |y^j - h(x^j)|$$

where $\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$ is a set of examples drawn from i.i.d. random variables with the same unknown $p(x, y)$. Let $h_m \in \operatorname{Arg\,min}_{h\in\mathcal{H}} R_{\mathcal{T}^m}(h)$ be a strategy with the minimal empirical risk.

a) What is the minimal $\varepsilon > 0$ which allows you to claim that the expected risk $R^{\text{MAE}}(h_m)$ is in the interval $(R_{\mathcal{T}^m}(h_m) - \varepsilon, R_{\mathcal{T}^m}(h_m) + \varepsilon)$ with probability 95% at least ?
b) What is the minimal number of the training examples m which guarantees that $R^{\text{MAE}}(h_m)$ is in the interval $(R_{\mathcal{T}^m}(h_m) - 1, R_{\mathcal{T}^m}(h_m) + 1)$ with probability 95% at least ?

Assignment 4. (*) Our task is to learn a prediction strategy $h\colon \mathcal{X} \to \{\text{male}, \text{female}\}$ estimating gender from a facial image $x \in \mathcal{X}$. We use our prior knowledge to design H different hypothesis spaces $\mathcal{H}_i \subset \mathcal{Y}^{\mathcal{X}}$, $i \in \{1, \ldots, H\}$. For example, each $\mathcal{H}_i$ can correspond to Convolutional Neural Networks with a different architecture. We randomly partition our i.i.d. drawn examples into three sets:

- $\mathcal{T}^m = \{(x^i, y^i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, m\}$ training set with m examples
- $\mathcal{V}^v = \{(x^i, y^i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, v\}$ validation set with v examples
- $\mathcal{S}^l = \{(x^i, y^i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, l\}$ test set with l examples

The prediction strategy is found in a two-stage process. In the first stage we apply ERM on the training set $\mathcal{T}^m$ to learn a strategy from each individual hypothesis space:

$$h_m^i \in \operatorname*{Arg\,min}_{h\in\mathcal{H}_i} R_{\mathcal{T}^m}(h)\,, \qquad i \in \{1, \ldots, H\}.$$

In the second stage, often called *model selection*, we apply the ERM on the validation set $\mathcal{V}^v$ to select the best hypothesis out of those learned in the first stage:

$$h_{\text{v}} \in \operatorname*{Arg\,min}_{i\in\{1,\ldots,H\}} R_{\mathcal{V}^v}(h_m^i)\ .$$

The very last step uses the test set $\mathcal{S}^l$ to evaluate the accuracy of the found hypothesis h_{v} by computing the test risk $R_{\mathcal{S}^l}(h_{\text{v}})$. In all cases the risks are computed using the 0/1-loss function $\ell(y, y') = [\![y \neq y']\!]$.

a) How would you chose the number of examples in the training, validation and the test set? *Hint: consider application of the solutions of Assignment 2 and 3.*

b) Assume that you applied the two-stage approach described above and evaluated the test risk of the found hypothesis. Let us consider three different results you could obtain:

	$R_{\mathcal{T}^m}(h_{\rm v})$	$R_{\mathcal{V}^v}(h_{\rm v})$	$R_{\mathcal{S}^l}(h_{\rm v})$
case 1	0.01%	14.2%	15.1%
case 2	3.6%	4.1%	12.3%
case 3	4.5%	4.8%	4.3%

What is the next reasonable step(s) you will take in order to improve the test accuracy? Consider each case separately. *Hint: your actions involve collecting new examples, changing the number of examples in trn/val/tst sets, using additional hypothesis spaces with higher/lower complexity etc.*

Assignment 5. (*) Our goal is estimate the expected risk $R^{0/1}(h) = \mathbb{E}_{(x,y)\sim p}[\![y \neq h(x)]\!]$ of a given prediction strategy $h\colon \mathcal{X} \to \{+1,-1\}$. To this end, we have collected independently two sets of examples. The first set $\mathcal{S}^{l_+} = \{x^i \in \mathcal{X} \mid i = 1,\ldots,l_+\}$ contains examples drawn i.i.d. from $p(x \mid y = +1)$, and the second set $\mathcal{S}^{l_-} = \{x^i \in \mathcal{X} \mid i = 1,\ldots,l_-\}$ examples drawn i.i.d. from $p(x \mid y = -1)$. Assume that we known the prior probability $p(y = +1)$. We estimate the value of $R^{0/1}(h)$ by computing

$$\hat{R}(h) = p(y = +1) \cdot \hat{R}_{\rm FN}(h) + p(y = -1) \cdot \hat{R}_{\rm FP}(h)\;, \tag{1}$$

where

$$\hat{R}_{\rm FN}(h) = \frac{1}{l_+} \sum_{x \in \mathcal{S}^{l_+}} [\![h(x) = -1]\!] \quad \text{and} \quad \hat{R}_{\rm FP}(h) = \frac{1}{l_+} \sum_{x \in \mathcal{S}^{l_-}} [\![h(x) = +1]\!]$$

is the empirical estimate of the false negative and the false positive rate, respectively.

a) Explain in what sense $\hat{R}(h)$ is a reasonable estimate of $R^{0/1}(h)$.

b) Find the smallest $\varepsilon > 0$ such that $R^{0/1}(h)$ is inside the interval $(\hat{R}(h) - \varepsilon, \hat{R}(h) + \varepsilon)$ with probability γ at least.

c) Evaluate ε for $\gamma = 0.95$, $p(y = +1) = 0.5$, $l_+ = 1000$ and $l_- = 20000$.