STATISTICAL MACHINE LEARNING (WS2016)
SEMINAR 1

Assignment 1. Consider the following parameter estimation task. You are given i.i.d. training data $\mathcal{T}^m = \{x_i \in \mathbb{R} \mid i = 1, 2, \ldots, m\}$ generated from the normal distribution

$$p_\mu(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

and the task is to estimate its unknown mean μ.

a) Show that the maximum likelihood estimator is given by the arithmetic mean of the training data, i.e.

$$\mu^* = e_{ML}(\mathcal{T}^m) = \frac{1}{m}\sum_{i=1}^{m} x_i .$$

Prove that this estimator is unbiased.

b) Compute the variance of the maximum likelihood estimator, i.e.

$$\mathbb{E}_\mu\big[(\mu - e_{ML}(\mathcal{T}^m))^2\big].$$

How does it depend on μ and m?

c) Someone (let us call him Mr. Y) proposes an even simpler estimator - he suggests to estimate the unknown mean μ of the distribution just by setting $\mu^* = x_1$ and dropping the rest of the training data. Moreover, he claims that his estimator is also unbiased. Show that this is true. Why would you nevertheless consider his estimator inferior compared to the maximum likelihood estimator?

Assignment 2. 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 3. We are given a prediction strategy with $h\colon \mathcal{X} \to \{-1, +1\}$. The task is to estimate the probability of misclassification $R^{0/1}(h) = \mathbb{E}_{(x,y)\sim p}(\llbracket y \neq h(x) \rrbracket)$ by computing the test error

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

where $\mathcal{S}^l = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, l\}$ is a set of examples drawn from i.i.d. random variables with the distribution $p(x, y)$. What is the minimal number of the test examples l which guarantees that the test error $R_{\mathcal{S}^l}(h)$ does not differ from $R^{0/1}(h)$ by more than $\varepsilon = 0.01$ (or 1%) with probability at least 95%?

Hint: Use the Hoeffding inequality which states that for all $\varepsilon > 0$ it holds that

$$\mathbb{P}\Big(\Big|\frac{1}{l}\sum_{i=1}^{l} z^i - \mu\Big| \geq \varepsilon\Big) \leq 2e^{-\frac{2l\,\varepsilon^2}{(b-a)^2}}$$

where $\{z^1, \ldots, z^l\} \in [a,b]^l$ are realizations of independent random variables with the same expected value μ.

Assignment 4. On the lecture we have shown the uniform convergence of the empirical risk $R_{\mathcal{T}^m}(h)$ to the expected risk $R(h)$ for a finite hypothesis space $\mathcal{H}$ by using the bound

$$\mathbb{P}\Big(\max_{h\in\mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon\Big) \leq 2|\mathcal{H}|e^{-\frac{2m\,\varepsilon^2}{(b-a)^2}}$$

where the risks are defined w.r.t. a loss function $\ell\colon \mathcal{Y}\times\mathcal{Y} \to [a,b]$. Use the same bound to prove the following theorem.

Theorem 1. Let $\mathcal{H}$ be a finite hypothesis space and $\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X}\times\mathcal{Y}) \mid i = 1,\ldots,m\}$ a training set drawn from i.i.d. random variables with distribution $p(x,y)$. Then, for any $\delta > 0$, any hypothesis $h \in \mathcal{H}$ and any loss function $\ell\colon \mathcal{Y}\times\mathcal{Y} \to [a,b]$, the inequality

$$R(h) \leq R_{\mathcal{T}^m}(h) + (b-a)\sqrt{\frac{\log 2|\mathcal{H}| + \log\frac{1}{\delta}}{2m}}$$

holds with probability at least $1-\delta$.

Assignment 5. [1] There have been 57 US presidential elections. There are $|\mathcal{H}| = 3100$ US counties. Let us see each county's voting outcome as a predictor $h\colon \mathcal{E} \to \mathcal{C}$, where $\mathcal{E} = \{1,2,\ldots\}$ is the set of election indices and $\mathcal{C}$ is a set of presidential candidates. Suppose there is a county h' which has always correctly predicted the elected US president. For example, $h'(1) =$ "George Washington", $h'(2) =$ "George Washington", ... , $h'(57) =$ "Barack Obama". What is the probability that this county will not predict the correct president in the future elections with confidence at least 95%?

Hint: Apply Theorem 1.

[1] Adopted from Xiaojin Zhu `http://pages.cs.wisc.edu/~jerryzhu/teaching.html`