Machine Intelligence 1 **Exercise Sheet 8**
WS 2022/23, Obermayer/Kashef/Strömsdörfer due: 12.01.2023 at 23:55

Statistical learning theory

Exercise T8.1: Empirical Risk Minimization (tutorial)

It is known from the literature[1] that the number of *affinely*[2] separable assignments of p data points in N dimensions is given by

$$\tilde{C}_{(p,N)} := 2\sum_{k=0}^{N} \left(\begin{array}{c} p-1 \\ k \end{array} \right).$$

(a) How is the Binomial coefficient $\binom{n}{k}$ defined for $n, k \in \mathbb{N}_0$ with $n \geq k$?

(b) Use the recursive formula for binomial coefficients

$$\left(\begin{array}{c} n \\ k \end{array} \right) + \left(\begin{array}{c} n \\ k-1 \end{array} \right) = \left(\begin{array}{c} n+1 \\ k \end{array} \right) \quad \text{to show that} \quad \tilde{C}_{(p,N)} + \tilde{C}_{(p,N-1)} = \tilde{C}_{(p+1,N)}\,.$$

Solution:

$$\begin{aligned} \tilde{C}_{(p,N)} + \tilde{C}_{(p,N-1)} &= 2\sum_{k=0}^{N}\binom{p-1}{k} + 2\sum_{k=0}^{N-1}\binom{p-1}{k} \\ &= \underbrace{2\binom{p-1}{0}}_{1=\binom{p}{0}} + 2\sum_{k=1}^{N}\Bigg[\underbrace{\binom{p-1}{k} + \binom{p-1}{k-1}}_{\binom{p}{k}}\Bigg] \quad = \quad \tilde{C}_{(p+1,N)}\,. \end{aligned}$$

(c) Explain how the number of affinely seperably assignments is related to the *Vapnik-Chervonenkis (VC) dimension* and overfitting.

Exercise H8.1: Vapnik-Chervonenkis dimension (homework, 2 points)

Use the definition of $\tilde{C}_{(p,N)}$ and the recursion property from above together with the binomial formula,

$$(x+y)^n = \sum_{k=0}^{n}\binom{n}{k} x^{n-k} y^k,$$

for $x, y \in \mathbb{R}$, $n \in \mathbb{N}$, to show that a linear classifier, $y(\underline{\mathbf{x}}; \underline{\mathbf{w}}) = \operatorname{sign}(w_0 + \sum_{i=1}^{N} x_i w_i)$, has a Vapnik-Chervonenkis dimension of $d_{\mathrm{VC}} = N + 1$.
Hint: It suffices to show that $\tilde{C}_{(N+1,N)} = 2^{N+1}$ and $\tilde{C}_{(N+2,N)} < 2^{N+2}$.

---

[1] Anthony, M., & Bartlett, P. L. (2009). Neural network learning: Theoretical foundations. cambridge university press. cf. Ch. 3.2 *The Growth Function*

[2] Note that in the lecture the formula for $C_{(p,N)}$ has the sum running to $N-1$ only (instead of N). This is due to the fact that the lecture considers *linear* separability in exactly N dimensions, i.e., no bias term, whereas here we consider separability using the linear *neuron* with bias, that is, $y(\underline{\mathbf{x}}) = \operatorname{sign}(w_0 + \sum_{i=1}^{N} w_i x_i)$. Therefore we define the tilde-variant $\tilde{C}_{(p,N)}$ here.

Exercise H8.2: Variability of classification (homework, 5 points)

Assume data $\underline{\mathbf{x}}^{(\alpha)} \in \mathbb{R}^N$ with $N = 2$ is drawn from two clusters C_1 and C_2 and distributed according to the following (multivariate) Normal distributions

$$\mathcal{N}(\underline{\boldsymbol{\mu}}_i, 2\,\underline{\mathbf{I}}), \qquad i = 1, 2$$

with $\underline{\boldsymbol{\mu}}_1 = (0,1)^\top$, $\underline{\boldsymbol{\mu}}_2 = (1,0)^\top$, and identity matrix $\underline{\mathbf{I}}$.

This task examines, how well a linear connectionist neuron can separate these two classes for increasing amounts p of available training data. Proceed as follows:

1. Generate a sample of $p/2$ data points $\underline{\mathbf{x}}^{(\alpha)}$ from each of the two clusters (i.e., p data points in total). Let $y^{(\alpha)} = 1$ for $\underline{\mathbf{x}}^{(\alpha)}$ from C_1 and $y^{(\alpha)} = -1$ for $\underline{\mathbf{x}}^{(\alpha)}$ from C_2.
2. Find the weights of a linear connectionist neuron with output

$$y(\underline{\mathbf{x}}; \underline{\mathbf{w}}) = \operatorname{sign}(w_0 + \sum_{i=1}^{N} w_i x_i)$$

   minimizing the squared error according to the known analytical formula for (see, e.g., problem T4.2c). There is no need to take the non-linearity $\operatorname{sign}(\cdot)$ into account.
3. Find the predictions of this classifier for $p_{\text{test}} = 1000$ new points drawn from the same distributions (again, 50% for each class).
4. Calculate the accuarcies (percentage of correct classifications) for the training (r_{train}) and test samples (r_{test}).

Repeat these steps 50 times for each $p \in \{3, 4, 6, 8, 10, 20, 40, 100\}$, and save the resulting parameters as well as the accuracies for training and testing.

Deliverables:

(a) (2 point) Plot the mean and standard deviation of r_{train}, and r_{test} against the number of training samples p in an errorbar-plot (plot p on the x-axis and the corresponding statistic on the y-axis of the figure).

(b) (1 point) Plot the means and standard deviation of w_1, w_2 and w_0 against p.

(c) (2 points) Interpret your results. How do these estimates depend on p?

Exercise H8.3: The Binomial distribution (homework, 3 points)

This exercise examines the relation between the following 3 distributions that are used in statistical learning theory:

$$f(k; n, p) = \binom{n}{k}\, p^k (1-p)^{n-k} \qquad \text{(Binomial distribution)}$$

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)} \qquad \text{(Normal distribution)}$$

$$f(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \qquad \text{(Poisson distribution)}$$

(a) (1 point) Visualize the probability mass function $f(k; n, p)$ of the binomial distribution for a few different values of k, n, p that demonstrate the different shapes that function can have.

(b) (1 point) The normal distribution is sometimes used as an approximation to the binomial distribution. Under which conditions is this reasonable? Under which conditions is it problematic? Visualize one example where the Normal approximation is good and one where it is not. Give at least one reason why this distribution is so widely used.

(c) (1 point) The Poisson distribution is often used as an alternative approximation to the binomial distribution. Under which conditions is it a good approximation? Visualize one example parametrization where the Poisson approximation is good and one where it is not.

Total 10 points.