Prof. Dr. M. Wolf
J. Cuesta

Mathematical Foundations of Machine Learning

Summer 2018
Blatt 6
(23.05.2018)

http://www-m5.ma.tum.de/Allgemeines/MA4801_2018S

The following homework will be discussed on May 29th and June 1st:

H6.1. (KL-Divergence) Let p and q be distributions on a finite alphabet $\mathcal{A}$. The Kullback-Leibler divergence, also known as the relative entropy, is defined in this case as

$$KL(p||q) := \sum_{a \in \mathcal{A}} p(a) \log \left[\frac{p(a)}{q(a)} \right],$$

if $q(a) = 0$ implies that $p(a) = 0$ for all $a \in \mathcal{A}$ (absolute continuity) and $KL(q||p) = \infty$ otherwise. Here we agreed that $0 \cdot \log \frac{0}{0} = 0$. Show that

a) $KL(p||q) \geq 0$ with equality only if $p = q$.

b) KL is jointly convex, i.e. for $0 \leq \lambda \leq 1$ and q_1, q_2, p_1, p_2 distributions on the finite alphabet $\mathcal{A}$,

$$KL(\lambda q_1 + (1-\lambda) q_2 || \lambda p_1 + (1-\lambda) p_2) \leq \lambda KL(q_1||p_1) + (1-\lambda) KL(q_2||p_2).$$

c) It satisfies the data-processing inequality: let q_X, p_X be distributions on $\mathcal{A}$ and $S(\cdot|\cdot)$ be a conditional probability such that

$$q_Y(\cdot) = \sum_{x \in \mathcal{A}} S(\cdot|x) q_X(x) \qquad \text{and} \qquad p_Y(\cdot) = \sum_{x \in \mathcal{A}} S(\cdot|x) p_X(x)$$

are distributions on a finite alphabet $\mathcal{B}$. Then

$$KL(q_Y||p_Y) \leq KL(q_X||p_X).$$

H6.2. Let P_{XY} be the joint distribution of two random variables X and Y whose marginals are respectively P_X and P_Y, then their mutual information is defined as

$$I(X:Y) := KL(P_{XY}||P_X \times P_Y).$$

a) Suppose that X and Y take values in the finite alphabet $\mathcal{A}$ and $\mathcal{B}$ respectively. Show that $I(X:Y) \leq \min(H(X), H(Y)) \leq \min(\log|\mathcal{A}|, \log|\mathcal{B}|)$ where $H(X) := -\sum_{a \in \mathcal{A}} p(a) \log p(a)$.

b) Show that if X, Y, Z are random variables that form a Markov chain $X - Y - Z$, then $I(X:Z) \leq I(X:Y)$.

c) Consider a stochastic learning algorithm, described by a probability measure μ_S on the function class $\mathcal{F}$, a risk function with values in $[0,1]$, and training data sets S drawn from a distribution over n independent elements. Prove that

$$\mathbb{E}_S \left[\hat{R}_S(\mu_S) - R(\mu_S) \right] \leq \sqrt{\frac{\log|\mathcal{F}|}{2n}}$$