Problem Set 8

For the Exercise Session on Dec 17

Last name	First name	SCIPER Nr	Points

Problem 1: Prediction and coding

After observing a binary sequence $u_1, \dots, u_i$, that contains $n_0(u^i)$ zeros and $n_1(u^i)$ ones, we are asked to estimate the probability that the next observation, u_{i+1} will be 0. One class of estimators are of the form

$$\hat{P}_{U_{i+1}|U^i}(0|u^i) = \frac{n_0(u^i)+\alpha}{n_0(u^i)+n_1(u^i)+2\alpha} \quad \hat{P}_{U_{i+1}|U^i}(1|u^i) = \frac{n_1(u^i)+\alpha}{n_0(u^i)+n_1(u^i)+2\alpha}.$$

We will consider the case $\alpha = 1/2$, this is known as the Krichevsky–Trofimov estimator. Note that for $i = 0$ we get $\hat{P}_{U_1}(0) = \hat{P}_{U_1}(1) = 1/2$.

Consider now the joint distribution $\hat{P}(u^n)$ on $\{0,1\}^n$ induced by this estimator,

$$\hat{P}(u^n) = \prod_{i=1}^{n} \hat{P}_{U_i|U^{i-1}}(u_i|u^{i-1}).$$

(a) Show, by induction on n that, for any n and any $u^n \in \{0,1\}^n$,

$$\hat{P}(u_1,\dots,u_n) \geq \frac{1}{2\sqrt{n}}\left(\frac{n_0}{n}\right)^{n_0}\left(\frac{n_1}{n}\right)^{n_1},$$

where $n_0 = n_0(u^n)$ and $n_1 = n_1(u^n)$.
[Hint: if $0 \leq m \leq n$, then $(1+1/n)^{n+1/2} \geq \frac{m+1}{m+1/2}(1+1/m)^m$]

(b) Conclude that there is a prefix-free code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ such that

$$\text{length}\,\mathcal{C}(u_1,\dots,u_n) \leq n h_2\left(\frac{n_0(u^n)}{n}\right) + \frac{1}{2}\log n + 2,$$

with $h_2(x) = -x\log x - (1-x)\log(1-x)$.

(c) Show that if $U_1, \dots, U_n$ are i.i.d. Bernoulli, then

$$\frac{1}{n}\mathbb{E}[\text{length}\,\mathcal{C}(U_1,\dots,U_n)] \leq H(U_1) + \frac{1}{2n}\log n + \frac{2}{n}$$

Problem 2: Lower bound on Expected Length

Suppose U is a random variable taking values in $\{1, 2, \dots\}$. Set $L = \lfloor \log_2 U \rfloor$. (I.e., $L = j$ if and only if $2^j \leq U < 2^{j+1}$; $j = 0, 1, 2, \dots$.

(a) Show that $H(U|L=j) \leq j$, $j = 0, 1, \ldots$.

(b) Show that $H(U|L) \leq \mathbb{E}[L]$.

(c) Show that $H(U) \leq \mathbb{E}[L] + H(L)$.

(d) Suppose that $\Pr(U=1) \geq \Pr(U=2) \geq \ldots$. Show that $1 \geq i \Pr(U=i)$.

(e) With U as in (d), and using the result of (d), show that $\mathbb{E}[\log_2 U] \leq H(U)$ and conclude that $\mathbb{E}[L] \leq H(U)$.

(f) Suppose that N is a random variable taking values in $\{0, 1, \ldots\}$ with distribution p_N and $\mathbb{E}[N] = \mu$. Let G be a geometric random variable with mean μ, i.e., $p_G(n) = \mu^n/(1+\mu)^{1+n}$, $n \geq 0$.

Show that $H(G) - H(N) = D(p_N \| p_G)$, and conclude that $H(N) \leq g(\mu)$ with $g(x) = (1+x)\log_2(1+x) - x\log_2 x$.

[Hint: Let $f(n,\mu) = -\log_2 p_G(n) = (n+1)\log_2(1+\mu) - n\log_2(\mu)$. First show that $\mathbb{E}[f(G,\mu)] = \mathbb{E}[f(N,\mu)]$, and consequently $H(G) = \sum_n p_N(n)\log_2(1/p_G(n))$.]

(g) Show that for U as in (d) and $g(x)$ as in (f),

$$E[L] \geq H(U) - g(H(U)).$$

[Hint: combine (f), (e), (c).]

(h) Now suppose U is a random variable taking values on an alphabet $\mathcal{U}$, and $c : \mathcal{U} \to \{0,1\}^*$ is an injective code. Show that

$$E[\text{length}\, c(U)] \geq H(U) - g(H(U)).$$

[Hint: the best injective code will label $\mathcal{U} = \{a_1, a_2, a_3, \ldots\}$ so that $\Pr(U = a_1) \geq \Pr(U = a_2) \geq \ldots$, and assign the binary sequences $\lambda, 0, 1, 00, 01, 10, 11, \ldots$ to the letters $a_1, a_2, \ldots$ in that order. Now observe that the i'th binary sequence in the list $\lambda, 0, 1, 00, 01, \ldots$ is of length $\lfloor \log_2 i \rfloor$.]

Problem 3: Tighter Generalization Bound

[10pts] Let $D = X_1, \ldots, X_n$ iid from an unknown distribution P_X, let $\mathcal{H}$ be a hypothesis space, and $\ell : \mathcal{H} \times \mathcal{X} \to \mathbb{R}$ be a σ^2-subgaussian loss function for every h. In the lecture we have seen that the generalization error can be upper bounded using the mutual information.

$$|\mathbb{E}_{P_{DH}}\left[L_{P_X}(H) - L_D(H)\right]| \leq \sqrt{\frac{2\sigma^2 I(D;H)}{n}}$$

(i) Modify the proof of the *Mutual Information Bound (11.2.2)* to show that if for all $h \in \mathcal{H}$, $\ell(h, X)$ is σ^2-subgaussian in X, then

$$|\mathbb{E}_{P_{DH}}\left[L_{P_X}(H) - L_D(H)\right]| \leq \sqrt{\frac{2\sigma^2 \sum_{i=1}^n I(X_i;H)}{n}}.$$

Hint: Recall from the lecture notes that

$$|\mathbb{E}_{P_{DH}}\left[L_{P_X}(H) - L_D(H)\right]| \leq \frac{1}{n}\sum_{i=1}^n \left|\mathbb{E}_{P_{X_i H}}\left[\ell(H, X_i)\right] - \mathbb{E}_{P_{X_i} P_H}\left[\ell(H, X_i)\right]\right|.$$

(ii) Show that, this new bound is never worse than the previous bound by showing that,

$$I(D;H) \geq \sum_{i=1}^n I(X_i;H).$$

(iii) Let us consider an example. Assume that $D = X_1, .., X_n,\ n > 1,$ are i.i.d. from $\mathcal{N}(\theta, 1)$, and that we do not know θ. We want to learn θ assuming the loss $\ell(h, x) = \min(1, (h - x)^2)$ (which is bounded) and $\mathcal{H} = \mathbb{R}$. Our learning algorithm outputs $H = \frac{1}{n}\sum_{i=1}^{n} X_i$. Use the new bound to show that

$$|\mathbb{E}_{P_{DH}}\left[L_{P_X}(H) - L_D(H)\right]| \leq \sqrt{\frac{1}{4(n-1)}}.$$

How does the old bound perform in this example?
Hint: Adding independent gaussian random variables, you get a gaussian random variable.