ECE 543 STATISTICAL LEARNING THEORY **SPRING 2019**
PROBLEM SET 6 **Due beginning of class, Thursday, April 25**

6. Minimax lower bounds using a bit of information theory

Assigned reading: Chapter 12 of the course notes.

1. **[Very noisy classification for a single feature]**
Consider a model free classification problem $(\mathsf{X}, \mathcal{F}, \mathsf{Y} = \{0,1\}, \mathcal{P})$ such that there is $x_0 \in \mathsf{X}$ such that $P\{X = x_o\} = 1$ for all $P \in \mathcal{P}$. The joint probability distributions of (X, Y) are thus parameterized by $p \in [0,1]$ such that $p = \mathbb{P}\{Y = 1\}$. Training data is given by $Z^n = (Z_1, \ldots, Z_n)$ such that $Z_i = (X_i, Y_i) = (x_o, Y_i)$, where $Y_1, \ldots, Y_n$ are independent Bernoulli(p) random variables. So Z^n has the same information as Y^n. An estimation algorithm A maps Y^n to $\{0,1\}$. The risk for a predictor $f \in \{0,1\}$ is $L_p(f) = \mathbb{P}_p\{Y \neq f\} = p(1-f) + f(1-p) = p + f(1-2p)$.

   (a) Find L_p^*, the minimum risk, and the optimal predictor, f^*, for a given p.

   (b) The expected excess risk for an algorithm A applied to n samples is $EL_p(\widehat{f}_n) - L_p^*$, where $\widehat{f}_n = A(Y^n)$ and the expectation is with respect to Y^n. Suppose $0 < h < 1$. Find a simple expression for the excess risk in case $p = \frac{1-h}{2}$ in terms of $E_0[\widehat{f}_n]$, where E_0 represents expectation assuming $Y_1, \ldots, Y_n$ are independent, Bernoulli($\frac{1-h}{2}$) random variables. (Hint: Your answer should reduce to zero if $E_0[\widehat{f}_n] = 0$.)

   (c) Find a simple expression for the excess risk in case $p = \frac{1+h}{2}$ in terms of $E_1[1 - \widehat{f}_n]$, where E_1 represents expectation assuming $Y_1, \ldots, Y_n$ are independent, Bernoulli($\frac{1+h}{2}$) random variables. (Hint: Your answer should reduce to zero if $E_1[1 - \widehat{f}_n] = 0$.)

   (d) Consider the binary hypothesis testing problem with hypotheses $H_0 : p = \frac{1-h}{2}$ vs. $H_1 : p = \frac{1+h}{2}$ and data Y^n. Let $p_e^*(h)$ denote the minimum average probability of error, assuming the hypotheses are equally likely. Explain why, for any learning algorithm, $\max_{p \in [0,1]} EL_p(\widehat{f}_n) - L_p^* \geq hp_e^*(h)$.

   (e) Find a constant C such that for any algorithm A, $\max_{p \in [0,1]} EL_p(\widehat{f}_n) - L_p^* \geq \frac{C}{\sqrt{n}}$. (Hint: Use part (d), the Bhattacharryya bound, and a suitable choice of h.)

2. **[Classification with many possible features, but limited concept class]**
Consider a model free classification problem $(\mathsf{X}, \mathcal{F}, \mathsf{Y} = \{0,1\}, \mathcal{P})$ such that X is a set with infinite cardinality and $\mathcal{F}$ is the family of all binary valued functions f on X such that the cardinality of $\{x : f(x) = 1\}$ is less than or equal to d, for some constant $d \geq 1$.

   (a) What is the VC dimension, V, of $\mathcal{F}$? What statistical guarantee on the performance of the ERM algorithm is implied by the fundamental theorem of learning theory (use the double version based on Dudley's bound).

   (b) Describe the ERM algorithm in detail. Be as explicit as possible. (Hint: Find the empirical loss for a given $f \in \mathcal{F}$ and Z^n, and try expressing it in such a way that it is easily minimized with respect to f subject to $f \in \mathcal{F}$.)

3. **[Classification with many possible features, limited concept class, and accurate labels]**
Consider a realizable classification problem $(\mathsf{X}, \mathcal{F}, \mathsf{Y} = \{0,1\}, \mathcal{P})$ (with no label noise) such that X is a set with infinite cardinality and $\mathcal{F}$ is the family of all binary valued functions f on X such that the cardinality of $\{x : f(x) = 1\}$ is less than or equal to d, for some constant $d \geq 1$. Consider the algorithm such that $\widehat{f}_n(x) = 1$ if $x \in \{X_i : 1 \leq i \leq n \text{ and } Y_i = 1\}$ and $\widehat{f}_n(x) = 0$ otherwise. Find the maximum, over all distributions P on X and all choices of $f \in \mathcal{F}$, of the expected excess risk for the algorithm. (Hint: The case $n + 1 \geq d$ is qualitatively different from $n + 1 < d$.)(Note: Your ERM algorithm in the previous problem should reduce to the algorithm mentioned here in the realizable case considered here.)

4. **[Concave-convex properties of mutual information]**
For simplicity, consider finite sets X and Y. A probability mass function (pmf) on X: p_X, and a conditional pmf, $p_{Y|X}$, determine a joint pmf on $\mathsf{X} \times \mathsf{Y}$ by $p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x)$. In turn, the joint pmf $p_{X,Y}$ determines the mutual information, $I(X;Y)$.

   (a) Show $I(X;Y)$ is a concave function of p_X (for $p_{Y|X}$ fixed). Hint: One approach is to use $I(X;Y) = \min_Q D(P_{Y|X}||Q|p_X)$, where $D(P_{Y|X}||Q|p_X) = \sum_i D(P_{Y|X=i}\|Q)p_X(i)$, for any pmf Q on Y.

   (b) Show $I(X;Y)$ is a convex function of $p_{Y|X}$ (for p_X fixed). Hint: One approach is to use the fact $I(X;Y) = D(p_{X,Y}\|p_X \otimes p_Y)$ and use the joint convexity of $D(p\|q)$ in (p,q).

5. **[Fano's inequality in information theory]**
Consider a joint distribution of (X,Y) such that X takes values in some set X with finite cardinality $|\mathsf{X}|$. Fano's inequality implies that if X can be accurately estimated from Y, then the conditional entropy, $H(X|Y)$, must be a small fraction of the maximum possible entropy of X, $\log|\mathsf{X}|$. Equivalently, $I(X;Y)$, equal to $H(X) - H(X|Y)$, must be close to the entropy of X. Let $\widehat{f}(Y)$ denote an estimator of X, let $B = \mathbf{1}_{\{X \neq \widehat{f}(Y)\}}$, and let $p_e = \mathbb{P}\{B = 1\}$. Explain why each of (a) - (d) below hold.

$$\begin{aligned}
H(X|Y) &\overset{(a)}{\leq} H(X,B|Y) \\
&\overset{(b}{=} H(B|Y) + H(X|B,Y) \\
&\overset{(c)}{\leq} H(B) + H(X|B=0,Y)(1-p_e) + H(X|B=1,Y)p_e \\
&\overset{(d)}{\leq} h(p_e) + p_e \log(|\mathsf{X}|-1)
\end{aligned}$$

Fano's inequality is $H(X|Y) \leq h(p_e) + p_e \log(|\mathsf{X}|-1)$, or, in weaker form, $H(X|Y) \leq \log 2 + p_e \log|\mathsf{X}|$.