# MATH 547: HOMEWORK 1
# DUE ON: FRIDAY, SEPTEMBER 15.

Please type the solutions in LaTeX, or write **very clearly** if you do it by hand. Lack of clarity in presentation and writing might result in a lower score.

**Problem 1, 15 points:**

Recall that the Bayes risk is the smallest possible risk of a binary classifier.

(1) The Bayes risk $L_*$ can take any value between $[0, 1]$ - true or false? Provide a short justification.

(2) Prove that the excess risk of any binary classifier $g$, defined as $\mathcal{E}(g) := L(g) - L_*$, can be expressed via

$$\mathcal{E}(g) = \int\limits_{\{x:g(x)\neq g_*(x)\}} |\eta(x)|\Pi(dx).$$

(3) If $g(x) = \text{sign}(\hat{\eta}(x))$ for some function $\hat{\eta} : S \mapsto \mathbb{R}$, then

$$\mathcal{E}(g) \le \int_S |\hat{\eta}(x) - \eta(x)|\Pi(dx) \le \left(\int_S |\hat{\eta}(x) - \eta(x)|^2\Pi(dx)\right)^{1/2}.$$

**Problem 2, 20 points (infinite base class):**

Assume that $X \in \mathbb{R}$ is a real-valued random variable, and

$$Y = 1 \text{ if } X \ge t_* \text{ and } Y = -1 \text{ if } X < t_*$$

for some fixed $t_* \in \mathbb{R}$ (that is, we are in the realizable learning framework). Let $G$ be the base class consisting of all binary classifiers of the form

$$g_t(x) = \begin{cases} +1, & x \ge t, \\ -1, & x < t. \end{cases} \text{ where } t \text{ can take any value in } \mathbb{R}.$$

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be the training data.

(1) Show that possible outputs of the empirical risk minimization (ERM) over class $G$ correspond to a set $\{g_t, t \in (a, b]\}$ for some $a, b \in \mathbb{R}$ that depend on the training data.

(2) Let $\varepsilon > 0$, and estimate $\Pr(L(\hat{g}) \ge \varepsilon)$ where $\hat{g}$ is any ERM solution of your choice.

(3) **(*)** Now describe an ERM solution $\hat{g}$ in the agnostic learning framework, that is, when there is no perfect classifier in $G$. Let $\bar{g}$ be the best classifier in $G$ (with the smallest generalization error) and estimate $\Pr(L(\hat{g}) - L(\bar{g}) \ge \varepsilon)$ in this case (for this last question only, assume that $|G| < \infty$, meaning that $G = \{g_{t_1}, \dots, g_{t_m}\}$ for some $m \ge 1$ and $g_{t_1}, \dots, g_{t_m} \in \mathbb{R}$).

**Problem 3, 20 points:**

(a) (A possible way to derive the kernel density estimator) Assume that $X$ is a real-valued random variable with distribution function $F$, and that $F$ is differentiable with $F' = p$. Suppose that $X_1, \dots, X_n$ are i.i.d. copies of $X$. Recall that the *empirical distribution function* is defined as

$$F_n(x) = \frac{1}{n}\sum_{j=1}^{n} I\{X_i \le x\}.$$

Then $F_n(x) \to F(x)$ almost surely for each $x$ [why?]
Now, assume that $F$ is 3 times continuously differentiable, and estimate the quality of approximation of $p$ by the "central difference" $\frac{F(x+h)-F(x-h)}{2h}$:

$$\left| p(x) - \frac{F(x+h)-F(x-h)}{2h} \right| \le ?$$

(here, $h > 0$ is a small positive constant).
Finally, show that replacing $F$ by $F_n$ in the central difference approximation of $p(x)$ yields the kernel density estimator (for a specific kernel function).

(b) Let $K(x) = I\{|x| \le 1/2\},\ x \in \mathbb{R}$. Let $(X,Y) \in \mathbb{R}\times\mathbb{R}$ be a random couple such that $(X,Y)$ has absolutely continuous distribution with the joint density $p_{X,Y}(x,y)$ and the marginal density $p_X(x)$ is known. Let $(X_1,Y_1),\dots,(X_n,Y_n)$ be i.i.d. copies of $(X,Y)$. Assume that the bandwidth parameter $h < \frac{1}{2}$ and derive the expression for the Nadaraya-Watson estimator in this case (we essentially did this in class, so just modify the argument carefully).

**Bonus problem, 20 points ("No Free Lunch" theorem):**

Assume that $X$ has uniform distribution on a finite set of points $S = \{x_1,\dots,x_M\}$, and that $Y = f_*(X)$ for some function $f_* : S \mapsto \{\pm 1\}$ (in other words, we are in the realizable learning framework). Let $\mathcal{X} = (X_1,Y_1),\dots,(X_n,Y_n)$ be the *training data*, an i.i.d. sample such that $n \le \lfloor M/2 \rfloor$ and $Y_j = f_*(X_j)$ for all $1 \le j \le n$. An *algorithm* $\mathcal{A}$ is any measurable mapping from $\mathcal{X}$ to the set of binary classifiers: in other words, $\mathcal{A}(\mathcal{X}) = \hat{f}_n$ where $\hat{f}_n$ is a binary classifier.

(a) Show that for any algorithm $\mathcal{A}$,

$$\max_{f_*:S\mapsto\{\pm1\}} \mathbb{E}_{\mathcal{X}}\Pr\Big(Y \ne \hat{f}_n(X)\,|\,\mathcal{X}\Big) = \mathbb{E}_{\mathcal{X}}\Pr\Big(f_*(X) \ne \hat{f}_n(X)\,|\,\mathcal{X}\Big) \ge \frac{1}{4}.$$

(you will still get full credit if you prove the bound with any positive constant rather than 1/4). Here, the expectation is taken with respect to the training data and $\Pr\Big(Y \ne \hat{f}_n(X)\,|\,\mathcal{X}\Big)$ denotes the conditional probability given the sample.
(hint: replace the maximum over $f_*$ by the average over all possible $f_*$. Then change the order of expectation/summation and consider what happens when $X \in \mathcal{X}$ and $X \notin \mathcal{X}$)

(b) Deduce from part (a) that

$$\Pr\Big(\Pr\Big(Y \ne \hat{f}_n(X)\,|\,\mathcal{X}\Big) \ge 1/8\Big) \ge 1/8.$$

In other words, if you are playing a game where player 1 picks any algorithm $\mathcal{A}$ and player 2 then picks the "problem" - the function $f_*$ that $\mathcal{A}$ has to learn, then player 2 can always guarantee that $\mathcal{A}$ fails with constant probability unless the training data covers most of the instances.