Homework Assignment 4

Due in class, Monday March 26th

SDS 384-11 Theoretical Statistics

1. (2 pts) Let $\mathcal{P}$ be the set of all distributions on the real line with finite first moment. Show that there does not exist a function $f(x)$ such that $Ef(X) = \mu^2$ for all $P \in \mathcal{P}$ where μ is the mean of P, and X is a random variable with distribution P.

   We must have $h(x)dP(x) = \mu^2$ for all distributions on the real line with mean μ. If P is degenerate at a point y, this implies that $h(y) = y^2$ for all y. But if P has mean zero ($\mu = 0$) and is not degenerate, then $h(x)dP(x) = x^2 dP(x) > 0 = \mu^2$. which is a contradiction.

2. (3 pts) A continuous distribution with CDF $F(x)$, on the real line is symmetric about the origin if, and only if, $1 - F(x) = F(-x)$ for all real x. This suggests using the parameter,

$$\theta(F) = \int (1 - F(x) - F(-x))^2 dF(x) \quad (1)$$

$$= \int ((1 - F(-x))^2 dF(x) - 2\int (1 - F(-x))F(x)dF(x) + \int F(x)^2 dF(x) \quad (2)$$

   as a nonparametric measure of how asymmetric the distribution is. Find a kernel h, of degree 3, such that $E_F h(X_1, X_2, X_3) = \theta(F)$ for all continuous F. Find the corresponding U statistic.

   Write for independent X_1, X_2, and X_3,

$$\theta(F) = \int P(X_1 > -x, X_2 > -x)dF(x) - 2\int P(X_1 > -x, X_2 < x)dF(x) + 1/3$$

$$= P(X_1 + X_3 > 0, X_2 + X_3 > 0) - 2P(X_1 + X_3 > 0, -X_2 + X_3 > 0) + 1/3$$

   This leads to the unbiased estimate of θ, $f(x_1, x_2, x_3) = I(x_1 + x_3 > 0, x_2 + x_3 > 0) - 2I(x_1 + x_3 > 0, -x_2 + x_3 > 0) + 1/3$. This is not symmetric in its arguments, so the symmetrized version has six terms, $h(x_1, x_2, x_3) = [f(x_1, x_2, x_3) + f(x_1, x_3, x_2) + f(x_2, x_1, x_3) + f(x_2, x_3, x_1) + f(x_3, x_1, x_2) + f(x_3, x_2, x_1)]/6$ The corresponding U-statistic is $U_n = \frac{1}{\binom{n}{3}} \sum_{i_1 < i_2 < i_3} h(X_{i_1}, X_{i_2}, X_{i_3})$.

   Many of you also expanded the last term out as $P(X_1 \leq X_3, X_2 \leq X_3)$. But note that since we have i.i.d random variables, this quantity is 1/3. I have given full score for this.

3. (3+3 pts)Look at the seminar paper "Probability Inequalities for Sums of Bounded Random Variables" by Wassily Hoeffding. It should be available via `lib.utexas.edu`. You can assume that n is a multiple of m (the degree of the kernel). Assume that the kernel is bounded, i.e. $|h(X_1, \ldots, X_m) - \theta| \leq b$, where $\theta = E[h(X_1, \ldots, X_m)]$.

(a) Read and reproduce the proof of equation 5.7 for large sample deviation of order m U statistics.

(b) Also prove Bernstein's inequality (see below) for U statistics. This is buried in the paper, you will have to find the bits and pieces and put them together. The Bernstein inequality is given by:

$$P(|U_n - \theta| \geq \epsilon) \leq a \exp\left(-\frac{n\epsilon^2/m}{c_1\sigma^2 + c_2\epsilon}\right),$$

where $\sigma^2 = \text{var}(h(X_1, \ldots, X_m))$ and c_1, c_2 are universal constants.

4. (3+3+3) Compute the VC dimension of the following function classes

(a) Circles in $\mathcal{R}^2$

For any three points which are not collinear , we can easily draw a circle that includes all three of them, any two of them, any one of them, or none of them, so VC dimension is at least 3.

However, for any set of four points, they are not shattered. We show this by constructing a counterexample in several cases:

i. Collinear: the labeling +-+- (going along the line) is impossible, among numerous others.

ii. Convex hull is a triangle: then the labeling with +(the three points of the triangle) and -(the interior point) is not possible.

iii. Convex hull is a quadrilateral: let (x_1, x_2) be the points separated by the long diagonal and (y_1, y_2) be the points separated by the short diagonal. At least one of the labelings $\{+x_1, +x_2, -y_1, -y_2\}$, $\{-x_1, -x_2, +y_1, +y_2\}$ will not be achieved. If they were both possible, then this would mean that the circles satisfying the two labelings can have four non-overlapping regions, which is not possible. (is it possible with ellipses?)Since some set of 3 points is shattered by the class of circles, and no set of 4 points is, the VC dimension of the class of circles is 3. Note that

(b) Axis aligned rectangles in $\mathcal{R}^2$ It is easy to see that one can shatter four points. Consider 5 points and the following cases.

i. They are all collinear. In which case, a trivial alternative labels cannot be shattered by axis aligned rectangles.

ii. If they are not all collinear, then draw the largest rectangle through the largest and smallest x and y coordinates. Either all five points are on this rectangle, or one is inside. In the first case, do an alternative labeling, this cannot be shattered by a axis aligned rectangle. In the second case, label everyone on the rectangle as one label, and the one inside with the opposite label.

(c) Axis aligned squares in $\mathcal{R}^2$ It is easy to construct 3 points which are shattered. Let us take 4 points. Let the leftmost point be L, rightmost R, top one T and bottom one B. Draw a rectangle like last question through these points. If there are ties, then it is easy to label them so that they cannot be shattered. So let us think about the case where there are no ties. In this case, if the rectangle is a square, i.e. the distance between x coordinates of L and R (d_x) is equal to the distance between y coordinates of T and B (d_y) are such that $d_x \geq d_y$, then (L+, R+, T-, B-) cannot be shattered. If $d_x < d_y$ then (L-, R-, T+, B+) cannot be shattered. So VC dimension is 4.