Homework 1
10/36-702
Due Friday February 3 at 3:00 pm

1 Mathematical Statistics warm-up (do you remember 705?)

(a) Suppose that $X_n \geq 0$ and $\mathbb{E}(X_n) = O(r_n)$. Prove that $X_n = O_{\mathbb{P}}(r_n)$.

(b) Suppose that $X_n \geq 0$ and $X_n = O_{\mathbb{P}}(r_n)$. Give an example to show that in general, this does not imply that $\mathbb{E}(X_n) = O(r_n)$.

(c) Prove that for $X \geq 0$, it holds that $\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > t)dt$. You may assume that X is continuously distributed and hence has a probability density function.

(d) Suppose that $X_n \geq 0$ and $X_n = O_{\mathbb{P}}(r_n)$, the latter bound holding "exponentially" fast, meaning that there are constants $\gamma_0, n_0 > 0$ such that for all $\gamma \geq \gamma_0$ and $n \geq n_0$, we have

$$X_n \leq \gamma r_n \text{ with probability at least } 1 - \exp(-\gamma).$$

Prove that $\mathbb{E}(X_n) = O(r_n)$. (Hint: use the formulation for $\mathbb{E}(X_n)$ from the last question.)

(e) Let $X_1, \ldots, X_n \sim P$, i.i.d., with $\mu = \mathbb{E}(X_i)$ and $\sigma^2 = \text{Var}(X_i)$. Define

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i, \quad S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

(i) Prove that $S_n^2 \xrightarrow{p} \sigma^2$.

(ii) Prove that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \xrightarrow{d} N(0,1).$$

(f) Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$.

(i) Prove that $\mathbb{E}(Y - f(X))^2$ is minimized by choosing $f^*(x) = \mathbb{E}(Y|X = x)$.

(ii) Prove that $\mathbb{E}(Y - X^T\beta)^2$ is minimized by choosing $\beta^* = B^{-1}\alpha$, where $B = \mathbb{E}(XX^T)$ and $\alpha = \mathbb{E}(YX)$.

2 Linear regression and linear classification

(a) Let $\Sigma = \mathbb{E}[XX^T]$. Assume that Σ is nonsingular. Let $v_1, \ldots, v_d$ be the (orthonormal) eigenvectors of Σ and let $\lambda_1, \ldots, \lambda_d$ be the corresponding eigenvectors. Let β^* minimize $\mathbb{E}(Y - X^T\beta)^2$.

(Below we will write $\langle u, v \rangle = u^T v$ for the inner product between vectors u, v.)

(i) Show that $\beta^* = \sum_{j=1}^d a_j v_j$ where

$$a_j = \frac{\mathbb{E}[\langle X, v_j \rangle Y]}{\lambda_j}, \quad j = 1, \ldots, d.$$

(ii) Show that, for any vector b,

$$\mathbb{E}[\langle X,b\rangle Y] = \mathbb{E}[\langle X,b\rangle \langle X,\beta^*\rangle],$$

and

$$\mathbb{E}[(Y - \langle b,X\rangle)^2] - \mathbb{E}[(Y - \langle \beta^*,X\rangle)^2] = \mathbb{E}[\langle b-\beta^*,X\rangle^2] = (b-\beta^*)^T\Sigma(b-\beta^*).$$

(iii) Suppose that the data are drawn independently $(X_1,Y_1),\ldots,(X_n,Y_n) \sim P$, where each $X_i \in \mathbb{R}$, $\mathbb{E}[X_i]=0$, and $\mathbb{E}[X_i^2]=1$. Also, assume that each Y_i and X_i are bounded. Let $(X,Y)\sim P$ be an independent pair. Let β^* minimize $\mathbb{E}(Y-X\beta)^2$ and let $\widehat{\beta}$ minimize $n^{-1}\sum_{i=1}^n (Y_i - \beta X_i)^2$. Show that

$$\int |x\widehat{\beta} - x\beta^*|^2 dP(x) = O_{\mathbb{P}}(1/n).$$

(b) **(Optional bonus problem)** For $\beta = (\beta_1,\beta_2,\ldots)$ a sequence of real numbers, let

$$|\beta_{(1)}| \geq |\beta_{(2)}| \geq |\beta_{(3)}| \geq \cdots$$

denote the components of β ordered by their decreasing absolute values. The weak ℓ_p norm of β, denoted by $\|\beta\|_{w,p}$, is the smallest C such that

$$|\beta_{(j)}| \leq \frac{C}{j^{1/p}}, \quad j = 1,2,\ldots$$

Let $\mathscr{D} = \{\psi_1,\psi_2,\ldots,\}$ be a countable collection of functions on $[0,1]$. Assume that $\int \psi_i(x)\psi_j(x)dx = 0$ for each $i \neq j$ and $\int \psi_j^2(x)dx = 1$ for each j. The weak L_p ball is defined as

$$L_{w,p}(C) = \left\{ f = \sum_j \beta_j \psi_j \,:\, \|\beta\|_{w,p} \leq C \right\}.$$

Define the best N-term approximation error

$$\sigma_N(f) = \inf_{|\Lambda|\leq N} \inf_{g\in \mathrm{Span}(\Lambda)} \|f-g\|,$$

where $\Lambda \subseteq \mathscr{D}$, Span$(\Lambda)$ denotes the set of linear combinations of functions in Λ, and $\|f-g\|^2 = \int (f(x)-g(x))^2 dx$. Show the following: if $f \in L_{w,p}(C)$ with $0<p<2$, then

$$\sigma_N(f) = O\left(\frac{1}{N^s}\right),$$

where $s = \frac{1}{p} - \frac{1}{2}$. Hint: an N-term approximation that achieves this order of error is in Span(Λ) where Λ contains the functions that correspond to $\beta_{(1)},\ldots,\beta_{(N)}$ in the expansion $f = \sum_j \beta_j \psi_j$.

(c) Suppose that $\mathbb{P}(Y=1) = \mathbb{P}(Y=-1) = \frac{1}{2}$ and further, the two class distributions are $X|Y=-1 \sim$ Uniform$(-10,5)$ and $X|Y=1 \sim$ Uniform$(-5,10)$. Derive an expression for the Bayes classifier, and the Bayes risk.

(d) Construct a concrete binary class classification example, in which the data from the two classes are linearly separable but the LDA solution does not separate the data. (A clear/thorough description is all that is needed, not a formal proof.)

3 Nonparametric regression

(a) Assume that we observe i.i.d. samples $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, $i = 1, \ldots, n$ from a model

$$y_i = f_0(x_i) + \epsilon_i, \quad i = 1, \ldots, n.$$

where ϵ_i, $i = 1, \ldots, n$ are independent with $\mathbb{E}(\epsilon_i) = 0$ and $\mathbb{E}(\epsilon_i^2) = \sigma^2$. For simplicity, we treat the input points $x_i \in [0,1]$, $i = 1, \ldots, n$ as fixed, satisfying the condition

$$P_n(I) \geq c|I|, \text{ for any interval } I \subseteq [0,1] \text{ with } |I| \geq 1/n,$$

where P_n is the empirical distribution of the input points. Also assume that $x \in [0,1]$. We also assume that the underlying regression function f_0 has a continuous, bounded derivative. Consider $\widehat{f}$, the kernel smoothing estimate with a boxcar kernel and bandwidth h.

(i) Prove that the squared bias and variance of $\widehat{f}$, at an arbitrary point x, satisfy

$$\underbrace{\left(\mathbb{E}[\widehat{f}(x)] - f_0(x)\right)^2}_{\text{Bias}^2(\widehat{f}(x))} \lesssim h^2 \quad \text{and} \quad \underbrace{\mathbb{E}\left[\left(\widehat{f}(x) - \mathbb{E}[\widehat{f}(x)]\right)^2\right]}_{\text{Var}(\widehat{f}(x))} \lesssim \frac{1}{nh}.$$

(Hint: use a Taylor expansion of f_0 around x.)

(ii) Derive the rate for the optimal choice of bandwidth h, and then give the error rate for the corresponding kernel smoothing estimator.

(b) Consider the univariate kth order local polynomial regression estimate, trained on the points $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, $i = 1, \ldots, n$, which we know can be expressed as

$$\widehat{f}(x) = \sum_{i=1}^{n} w_i(x) y_i,$$

for some weights $w_i(x)$, $i = 1, \ldots n$.

(i) Prove that at any point $x \in \mathbb{R}$, we have

$$\sum_{i=1}^{n} w_i(x) = 1 \quad \text{and} \quad \sum_{i=1}^{n} w_i(x)(x_i - x)^j = 0, \text{ for } j = 1, \ldots k.$$

(ii) Now assume the same model for the data as in part (a), and further assume that $x_i = i/n$, $i = 1, \ldots, n$, and $f_0 \in H_1(k+1, L)$ for a positive integer k and a constant $L > 0$. Also assume that $x \in [0,1]$. Take $\widehat{f}$ to be the local polynomial regression estimate of order k, and compute the bias of $\widehat{f}$ at an arbitrary point x, using the results in the last part. (Hint: use a Taylor expansion, and the results of the last question. You can use the fact that $\sum_{i=1}^{n} |w_i(x)| \leq C$ for some constant C that does not depend on n or h.) What do you conclude about the bias of local polynomial regression, compared that of kernel regression?

(iii) Why don't we just keep increasing the polynomial order k without end? You can answer this either with some theory, or a quick simulation. (Hint: consider the variance of $\widehat{f}$. You can use the fact that $\sum_{i=1}^{n} w_i^2(x)$ is an increasing function of k.)

(c) **(Optional bonus problem)** Suppose that, in backfitting, we choose our univariate smoother just to be standard univariate linear regression. Prove that backfitting converges in one pass, and the resulting estimate is just standard multivariate linear regression.