Stats315a Problem Set 1
Due: Friday, October 8 by 11:59pm on Gradescope.

Question 1.1 (Ridge regression risks, 20pts)**:** Consider the ℓ_2-regularized (ridge) regression estimator

$$\widehat{\beta}_\lambda := \operatorname*{argmin}_b \left\{ \frac{1}{2} \|Xb - y\|_2^2 + \frac{\lambda}{2} \|b\|_2^2 \right\},$$

where $X = [x_1 \ \cdots \ x_n]^T \in \mathbb{R}^{n \times p}$ is the (fixed) design matrix and $y \in \mathbb{R}^n$ is the response. Let $H_\lambda = X(X^TX + \lambda I)^{-1}X^T$, and assume that

$$y_i = f(x_i) + \varepsilon_i \tag{1.1}$$

where $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] = \sigma^2$. (Note that we *do not* assume that $y_i = x^T\beta^\star + \varepsilon_i$.) Recall also that the in-sample risk of an estimate $\widehat{f}$ of f is

$$R_{\text{in}}(\widehat{f}) := \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(\widehat{f}(x_i) - f(x_i))^2] = \frac{1}{n}\sum_{i=1}^n \left(\text{Bias}(\widehat{f}(x_i))^2 + \text{Var}(\widehat{f}(x_i)) \right)$$

where the expectation is taken over y_i drawn in model (1.1). Define the mean values of y by

$$\mu := \mathbb{E}[y] = [f(x_i)]_{i=1}^n.$$

Throughout the remainder of the question, let $\widehat{f}_\lambda$ be the linear function $\widehat{f}_\lambda(x) = x^T\widehat{\beta}_\lambda$ given by the ridge estimator.

(a) Show that

$$n \cdot R_{\text{in}}(\widehat{f}_\lambda) = \|(I - H_\lambda)\mu\|_2^2 + \sigma^2 \operatorname{tr}(H_\lambda^T H_\lambda).$$

(b) Show that the residual sum of squares $\text{RSS} = \sum_{i=1}^n (\widehat{f}_\lambda(x_i) - y_i)^2$ (this is just the training squared error) satisfies

$$\mathbb{E}[\text{RSS}] = nR_{\text{in}}(\widehat{f}_\lambda) + \sigma^2(n - 2\operatorname{tr}(H_\lambda)).$$

For the remainder of the question, assume that the design $X \in \mathbb{R}^{n \times p}$ has rank p, that is, it is full column rank.

(c) Let $X = U\Gamma V^T$ be the singular value decomposition (SVD) of X, where $U \in \mathbb{R}^{n \times p}$ satisfies $U^TU = I_p$ and $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_p)$ is the diagonal matrix of singular values. Using this SVD, give as explicit a formula as you can for the derivative matrix

$$\dot{H}_\lambda := \frac{\partial}{\partial \lambda} H_\lambda \in \mathbb{R}^{n \times n}.$$

(d) Let $r(\lambda) = n \cdot R_{\text{in}}(\widehat{f}_\lambda)$ be the in-sample risk as a function of $\lambda \geq 0$. Give a formula for the derivative $r'(\lambda) = \frac{\partial}{\partial \lambda} r(\lambda)$.

(e) Using your preceding two answers, show that $r'(0) < 0$, that is, there is *always* some $\lambda > 0$ so that the in-sample risk of the ridge estimator is smaller than unregularized least squares.

Question 1.2 (The choice of loss functions, 20pts)**:** Consider the margin-based classification problem with data in pairs $(x, y) \in \mathcal{X} \times \{-1, 1\}$, where we seek a predictor $f : \mathcal{X} \to \mathbb{R}$ with large margin $yf(x)$. Let the loss

$$\ell(s, y) = \phi(sy)$$

for a convex $\phi : \mathbb{R} \to \mathbb{R}_+$. The loss is *infinite sample consistent* (or just consistent) for the zero-one error if for any distribution on Y, where $p = \mathbb{P}(Y = 1)$, the minimizer

$$s_\phi^\star(p) := \operatorname*{argmin}_{s \in \mathbb{R}} \{\mathbb{E}[\ell(s, Y)] = p\phi(s) + (1 - p)\phi(-s)\}$$

satisfies

$$\operatorname{sign}(s_\phi^\star(p)) = \operatorname{sign}(2p - 1)$$

whenever $p \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$. In the case that $p \in \{0, 1\}$, we require that

$$\inf_{s(2p-1) \le 0} \{p\phi(s) + (1 - p)\phi(-s)\} > \inf_{s \in \mathbb{R}} \{p\phi(s) + (1 - p)\phi(-s)\},$$

but we will ignore that for this question.

(a) Show that if ϕ is differentiable and $\phi'(0) < 0$, then the loss is consistent.

(b) Let ϕ be differentiable with $\phi'(0) < 0$ and $\lim_{s\to\infty} \phi(s) = 0$. Give a transformation $h : \mathbb{R} \to [0, 1]$ from scores s to probabilities such that

$$s^\star = \operatorname{argmin} \{p\phi(s) + (1 - p)\phi(-s)\} \quad \text{if and only if} \quad p = h(s^\star).$$

Hint. You may use that for a convex ϕ, the derivative ϕ' is non-decreasing.

(c) Let $\phi_{\log}(s) = \log(1 + e^{-s})$ be the logistic loss. Give $s_\phi^\star(p)$ and the transformation h.

(d) Let $\phi_{\exp}(s) = \exp(-s)$ be the exponential loss. Give $s_\phi^\star(p)$ and the transformation h.

(e) Let ϕ be the hinge loss $\phi(s) = (1 - s)_+$. Give $s_\phi^\star(p)$, and show that there is no transformation of the form in part (b).

Question 1.3 (The curse of dimensionality and distances in high dimensions, 15pts)**:** Let $X_i \in \{-1, 1\}^p$ be uniformly distributed on the hypercube, and let P be the uniform distribution on $\{-1, 1\}^p$, so that $X_i \overset{\text{iid}}{\sim} P$.

(a) Show that for any vector $v \in \mathbb{R}^p$,

$$\mathbb{E}[\exp(X^T v)] \le \exp\left(\frac{1}{2} \|v\|_2^2\right).$$

It may be useful to use that $\frac{1}{2}(e^t + e^{-t}) \le e^{t^2/2}$, valid for all $t \in \mathbb{R}$.

(b) Show that for any independent $X_1, X_2 \overset{\text{iid}}{\sim} P$ and $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp(\lambda X_1^T X_2)] \le \exp\left(\frac{\lambda^2 p}{2}\right).$$

(c) Using a Chernoff bound, show that for any $t \geq 0$,

$$\mathbb{P}\left(\|X_1 - X_2\|_2^2 \leq 2p(1-t)\right) \leq \exp\left(-\frac{pt^2}{2}\right) \text{ and } \mathbb{P}\left(\|X_1 - X_2\|_2^2 \geq 2p(1+t)\right) \leq \exp\left(-\frac{pt^2}{2}\right).$$

(d) Let $X_i \overset{\text{iid}}{\sim} P$ for $i = 1, \ldots, N$ and $\delta \in (0,1)$. Show that if

$$N \leq \exp\left(\frac{pt^2}{4} - \frac{1}{2}\log\frac{1}{\delta}\right)$$

then $2p(1-t) \leq \|X_i - X_j\|_2^2 \leq 2p(1+t)$ for all $i \neq j$ with probability at least $1 - \delta$.

(e) Conclude that even if we draw a sample of size N exponential in the dimension p, we expect each pair X_i, X_j, $i \neq j$, to have ℓ_2 distance $\|X_i - X_j\|_2 \approx \sqrt{2p}$ with high probability.

Question 1.4 (Limiting ridge solutions, 10pts)**:** Let $\widehat{\beta}_\lambda = \text{argmin}_b\{\|Xb - y\|_2^2 + \lambda \|b\|_2^2\}$ be the ridge regression estimator. Let $X \in \mathbb{R}^{n \times p}$ and assume $p > n$, where X has rank n. Using the SVD of X, give a closed form for $\lim_{\lambda \downarrow 0} \widehat{\beta}_\lambda$.

Question 1.5 (Linear regression versus k-nearest neighbors, 30pts)**:** You compare k-nearest neighbors (knn) and linear regression in terms of their classification performance in the presence of increasing numbers of noise variables. The setup is as follows, and mimics the mixture simulation in class. The 20×2 data matrix `mixturemeans.csv` is available on the course website; the first 10 rows are for class 1, the next 10 for class 2. Let $M_1 = [\mu_1 \ \cdots \ \mu_{10}]^T \in \mathbb{R}^{10 \times 2}$ and $M_2 = [\mu_{11} \ \cdots \ \mu_{20}]^T \in \mathbb{R}^{10 \times 2}$ be these matrices of means, where $\mu_i \in \mathbb{R}^2$.

(a) Write a function to generate a sample of N points from a uniform mixture of Gaussians in $\mathbb{R}^2$, with each Gaussian $\mathsf{N}(\mu_i, \sigma^2 I)$ having diagonal covariance $\sigma^2 I$ for a fixed $\sigma^2 > 0$. The function takes as inputs the centroid matrix M, sample size N and σ, and outputs a matrix $X \in \mathbb{R}^{N \times p}$ whose rows are i.i.d. draws from this mixture of Gaussians.

(b) Use your function to generate a dataset of size $N_{\text{train}} = 100$ for each of the two classes, with $\sigma^2 = \frac{1}{5}$, as well as a test set of size $N_{\text{test}} = 10^4$ for each class. Create the corresponding response vectors for each. This should leave you with matrices $X_{\text{train}} \in \mathbb{R}^{2N \times 2}, X_{\text{test}} \in \mathbb{R}^{2N_{\text{test}} \times 2}$ and responses y_{train} and y_{test}.

(c) What is the Bayes (optimal) classifier for this problem? Write this in terms of the densities $f_i, i = 1, \ldots, 20$ for each of the mixture components.

(d) Write a function to compute the Bayes classifier for this setup. It should take as input the two matrices M_0, M_1 of means, variance σ^2, and an input matrix X to be classified. Your function should classifiy all the rows of X.

(e) Write an evaluation function that takes as input your training data, test data, and a vector of values for k, the knn neighborhood size parameter. Your function should

i. Estimate the Bayes error using the test data using your function from part (d).

ii. Estimate the test error of a linear classifier fit by least squares.

iii. Estimate the test errors for knn at each of the values of k (in `R`, the package `class` has a `knn` function).

Run your function using $k = 1, 3, 5, 7, 9, 11, 13, 15$.

(f) Write a new function that expands the evaluation function in the part (e) to take two extra parameters: the number noise of noise variables and a variance τ^2_{noise}. This function adds additional Gaussian noise columns to X_{train} and X_{test}, where the noise columns have i.i.d. $\mathsf{N}(0, \tau^2_{\text{noise}})$ entries. This function should produce the same outputs as that in part (e). Run your function with $p_{\text{noise}} = 1, 2, \ldots, 10$ noise variables with $\tau^2_{\text{noise}} = 1$. Summarize its outputs.

Question 1.6 (Causal estimation, 10pts)**:** Consider the potential outcomes framework for a real-valued response Y with randomized treatment assignments $W \in \{0, 1\}$, so that $(Y(0), Y(1)) \perp W$. Let

$$\tau^\star := \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

be the average treatment effect, which is the variable of interest. Assume there are covariates $X \in \mathbb{R}^p$, which may (or may not) be related to the responses Y, but where W is also independent of X. Let the population mean-square-error estimates be

$$(\tau_{\text{mse}}, \alpha_{\text{mse}}, \beta_{\text{mse}}) = \operatorname*{argmin}_{\tau,\alpha,\beta} \mathbb{E}\left[(Y - \alpha - X^T\beta - \tau W)^2\right],$$

so that $\alpha_{\text{mse}} \in \mathbb{R}$ is an intercept, $\beta_{\text{mse}} \in \mathbb{R}^p$, and τ_{mse} is the coefficient of W in the model $Y_i = \alpha + X_i^T\beta + \tau W_i + \varepsilon_i$. Show that

$$\tau_{\text{mse}} = \tau^\star.$$