Yale University
Department of Statistics and Data Science
Midterm

STATISTICS 365/565

Issued: 04/22/2021 **Due:** 04/24/2021

Notes: You will have three hours. You cannot discuss this exam with anybody at any time before 04/24/2021 (inclusive). You *can* use notes, online resource, videos, etc... Just nothing adaptive on which you can ask a direct question and get it answered (e.g. no stackoverflow/slack/asking a friend/etc...).

Submission: You will submit this to gradescope as a PDF.

Problem 1: Cross-validation Consider a collection of data points $\{x_i, y_i\}_{i=1}^n$ and a learning algorithm that depends on some parameter λ and outputs some function f_λ that maps examples x to predicted labels y. Suppose that the loss we are interested in for some new example x, is $(f(x) - y)^2$ where y is the true label. How would you use cross-validation to pick a model with a good λ that will perform well on new (unseen) data-points?

Problem 2: Non-linear embeddings and gradients One approach to perform non-linear embeddings of data is to use neural networks. Let $g(s) : \mathbb{R} \mapsto \mathbb{R}$ be the ReLU function. We assume that if we apply g to a vector $v \in \mathbb{R}^d$, then $g(v)$ is applied elementwise. Suppose that we have a collection of data points $x_i \in \mathbb{R}^d$. Consider the following operations

$$a^2 = g(W^1 x)$$
$$\tilde{x} = W^2(a^2)$$

where $a^2 \in \mathbb{R}^k$, $W^1 \in \mathbb{R}^{k\times d}$, and $W^2 \in \mathbb{R}^{d\times k}$. For $k < d$ we treat a^2 as a low-dimensional embedding of our data point x, and we take $\tilde{x}$ to be the reconstructed version of x. Therefore, we would want $\tilde{x}$ to be close to x.

To that end consider the following optimization

$$\arg\min_{W^1, W^2} \sum_{i=1}^n \|W^2(g(W^1 x_i)) - x_i\|^2$$

In a few sentences describe why we might expect a^2 is a good low-dimensional embedding of the data point x. Devise the stochastic gradient descent update for the above problem.

Problem 3: Boosting In class we saw boosting. One specific version was adaboost. We saw that that was a special case for a more general procedure using forward step-wise regression where we keep adding a function to minimize the objective. We also saw gradient boosting.

Devise a boosting strategy to optimize the following objective

$$\arg\min_f \sum_{i=1}^n |f(x_i) - y_i|$$

Assume that you have access to a tree learning procedure that can build any depth seven tree and can optimize any objective that you need it to.

Problem 4: Clustering with least absolute deviation Consider the setting where we have $x_i \in \mathbb{R}$. We wish to find a collection of cluster representatives $\mu_1, \mu_2, \ldots, \mu_K$ and an assignment $\pi(i) \in [K]$ that assigns example i to one of the K representatives. We wish to minimize

$$\arg\min_{\pi,\mu_j} \sum_{i=1}^{n} |x_i - \mu_{\pi(i)}|$$

The above optimization is computationally intractable in general. Devise an iterative algorithm similar to K–means to solve the above problem.

Problem 5: PCA Suppose that we are given data $X \in \mathbb{R}^{p\times n}$ where the columns of X represent our examples $x_i \in \mathbb{R}^p$. Recall the SVD of $X = USV^T$. We wish to find a low-dimensional embedding $\alpha_i = Ax_i$ where $A \in \mathbb{R}^{k\times p}$ and the columns are orthogonal. We wish to maximize $\sum_i \|\alpha_i\|_2^2$. What is α_i as a function of S and V?

Problem 6: Lost your data Suppose that you are given a matrix $M \in \mathbb{R}^{n\times n}$, such that $M_{(ij)} = \langle x_i, x_j\rangle$, for your data $x_i \in \mathbb{R}^p$. Suppose that you do not actually have the data. This often happens if the data has large, possibly infinite, dimensions. Another instance is if the data itself cannot be represented as a straight-forward feature vector. For instance similarity metrics between proteins exist, but a protein just exists in a very high-dimensional space.

We wish to find the embedding vectors $\alpha_i \in \mathbb{R}^k$ such that

$$\sum_{i,j} (\alpha_i^T \alpha_j - M_{(ij)})^2$$

is minimized.

Suppose that $X \in \mathbb{R}^{p\times n}$ is your data matrix. Then $M = X^T X$. Use the SVD of X to find the SVD of M. Then use that to compute the embeddings.

Problem 7: Regularization and SVD

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\|X\beta - y\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2$$

Let $X = USV^T$ and $k = \text{rank}(X)$. Show that $\widehat{\beta} = V_k(S_k^2 + \lambda I_k)^{-1} S_k U_k^T y$ is a viable solution. Also show that $\widehat{\beta} = \sum_{j=1}^{k} \frac{\sigma_j}{\sigma_j^2+\lambda} v_j u_j^T y$.

Problem 8: Low-rank regression, or regression with dimensionality reduction Suppose we have data (x_i, y_i) for regression. It is common practice to perform dimensionality reduction on the data before doing the regression if the data is very high-dimensional. Doing so also serves as a denoising operation. Let $X = USV^T$ be the SVD of X.

- Let $A \in \mathbb{R}^{n\times k}$ be the matrix whose rows are the α_i. Perform least squares on the objective $\|A\beta_{lr} - y\|_2^2$. What is β_{lr}? Your solution should be in terms of U_k, S_k, V_k, and y. (lr stands for low-rank here)

Problem 9: Over-complete least squares Suppose we have a data matrix $X \in \mathbb{R}^{n \times p}$ where $p > n$, and the rank of X is n. We wish to perform linear regression to optimize

$$\arg\min_{\beta} \|X\beta - y\|_2^2$$

However, X has a null-space is $p > n$. That is there exist vectors $v \in \mathbb{R}^p$ such that $Xv = 0$. As a result, the typical solution $\widehat{\beta} = (X^T X)^{-1} X^T y$ is not valid since $X^T X$ is not invertible. Nevertheless, the gradient optimality condition still applies $X^T(X\widehat{\beta} - y) = 0$. Using the SVD of X and the fact that many of the singular values of X are zero, find a potential solution for $\widehat{\beta}$.

Problem 10: Extending previous problem Now find a solution such that the ℓ_2 norm of $\widehat{\beta}$ is minimized over all possible choices. That is, if $V = \{\beta \mid X^T X\beta = X^T y\}$ is the set of all β that satisfy the optimality conditions, then find

$$\widehat{\beta} = \arg\min_{\beta \in V} \|\beta\|_2$$

Note that this is very related to the projections we have discussed in class.