Yale University
Department of Statistics and Data Science
Midterm

STATISTICS 365/565

Issued: 03/15/2021 **Due:** 03/17/2021

Notes: You will have three hours. You cannot discuss this exam with anybody at any time before 03/17/2021 (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: Gradients Consider the following:

$$g(A, \beta, x) = \sum_{i=1}^{p} \beta_{(i)} \left(\sum_{j=1}^{k} A_{(ij)} x_{(j)} \right)^3$$

where $\beta \in \mathbb{R}^p$, $A \in \mathbb{R}^{p \times k}$ and $x \in \mathbb{R}^k$ Compute

$$\nabla_A g$$
$$\nabla_x g$$
$$\nabla_\beta g$$

Please write these in matrix notation. Recall the Hadamard (pointwise) product between two vectors is $z = (v \circ w)$ is the vector such that $z_{(i)} = v_{(i)} w_{(i)}$. Furthermore, for a vector v you may denote $z = v^3$ as the vector such that $z_{(i)} = v_{(i)}^3$. Thus, in matrix notation the above is equal

$$g(A, \beta, x) = \beta^T (Ax)^3$$

Problem 2: Weighted logistic regression Consider a binary classification problem with $y \in \{0, 1\}$. At times our data might be very imbalanced towards one label, for instance it might be difficult to obtain $y = 1$ samples. In such situations we may need to weight samples differently. Consider the following weighted logistic regression loss:

$$L(\theta) = \sum_{i=1}^{n} d_i \left(\log(1 + \exp(x_i^T \theta)) - y_i x_i^T \theta \right)$$

where $d_i \in \mathbb{R}$ and $d_i \geq 0$. What is $\nabla L(\theta)$? Your solution does not have to be in matrix notation. For instance, you can simply specify the i^{th} coordinate of the gradient and can also leave the solution with a summation.

Problem 3: Weighted linear regression At times we might have a belief that some of our data samples have different noise variances. Consider the following model

$$y_i = x_i^T \beta^* + w_i$$

where $w_i \sim N(0, \sigma_i^2)$ are independent. Suppose that we observe $\{(x_i, y_i, \sigma_i^2)\}_{i=1}^n$. What is the negative log-likelihood for this problem? How does this change over standard linear regression?

Problem 4: Regularization Suppose that we have the following optimization problem

$$\arg\min_{\theta} f(\theta) + \frac{\lambda}{2}\|\theta\|_2^2$$

What is the gradient descent update for this problem? Assume that f is differentiable. Don't forget the learning rate. Your answer should take the form

$$\theta_{k+1} = \theta_k - \text{stuff goes here}$$

Problem 5: Concepts The following question has to do with KNN classification.

- a) Say we have collected 100 flowers each belonging species A or species B. For each flower we measure the stem length, petal diameter, petal width, and sepal length and also note what species the flower is (species A or B). Explain in words how the nearest neighbor classifier for the species based on the other measurements with $k = 5$ would classify a new test case.
- b) Why would it be a bad idea to use $k = 1$ in our classifier? How might this affect the test error?
- c) Why would it be a bad idea to use $k = 100$ in our classifier?

Problem 6: Gradient Descent Suppose that we define the following loss

$$\ell(y, y') = \begin{cases} \frac{1}{2}(y - y')^2 & \text{if } |y - y'| < 1 \\ |y - y'| - \frac{1}{2} & \text{otherwise} \end{cases}$$

Suppose that we have n data points of the form $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We wish to solve the optimization

$$\operatorname*{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell(x_i^T \theta, y_i)$$

One approach to solving the above optimization is to use gradient descent. To use gradient descent we would need to compute the gradient of $\ell(x_i^T \theta, y_i)$ where the gradient is taken with respect to the parameters θ. What is $\nabla_\theta \ell(x_i^T \theta, y_i)$? That is the gradient for the i^{th} example.

Problem 7: MLE Suppose you observe n data $y_i = x_i^T \theta^* + w_i$ where all w_i are independent and follow an exponential PDF with $f_w(w) = \exp(-w)\mathbb{1}(w \geq 0)$. What is the log-likelihood for some parameter θ?

Problem 8: MLE Suppose you observe n i.i.d. data $x_i \sim N(0, \sigma^2)$. What is the MLE estimate for σ^2 where you know that the mean is 0?

Problem 9: MLE Suppose you observe n i.i.d. data $y_i \in \{0,1\}^k$. What this means is that the entries of y_i are either 0 or 1. Furthermore, we assume that only one entry of y_i is equal to one and $\mathbb{P}([y_i]_{(j)} = 1) = p_j$ where $p_j \geq 0$ and $\sum_{j=1}^k p_j = 1$. Suppose that $p_j = \exp(\theta_j)/\sum_{j=1}^k \exp(\theta_j)$. What is the MLE estimate for θ_j? Note that it is not necessarily unique! To simplify your life assume that your estimates satisfy $\sum_{j=1}^k \exp(\widehat{\theta}_j) = 1$. Then find your estimates $\widehat{\theta}_j$ of θ_j, and verify that that it is the case that $\sum_{j=1}^k \exp(\widehat{\theta}_j) = 1$. You may also assume that for each of $j \in [k]$ you have at least one observation such that $(y_i)_{[j]} = 1$.

Problem 10: Concept Suppose you trained an NN classifier on data $\{((x_i - m)/s, y_i)\}_{i=1}^n$ with $y_i \in \{0,1\}$. Now you have a function h that maps the input to 0 or 1. Suppose that you observe a new data point x. How would you create your prediction for the label of x?