Math 583 Exam 1 is on Friday, Sept. 22 and covers homeworks 1-3 and quizzes 1-3. You are allowed 7 sheets of notes and a calculator. Any needed tables will be provided. CHECK FORMULAS: YOU ARE RESPONSIBLE FOR ANY ERRORS ON THIS HANDOUT!

1) Statistical Learning techniques extract information from multivariate data. A **case** or **observation** consists of k random variables measured for one person or thing. The ith case $\boldsymbol{z}_i = (z_{i1}, ..., z_{ik})^T$. The **training data** consists of $\boldsymbol{z}_1, ..., \boldsymbol{z}_n$. A statistical model or method is fit (trained) on the training data. The **test data** consists of $\boldsymbol{z}_{n+1}, ..., \boldsymbol{z}_{n+m}$, and the test data is often used to evaluate the quality of the fitted model.

2) The focus of *supervised learning* is predicting a future value of the response variable Y_f given $\boldsymbol{x}_f$ and the training data $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$. The focus of *unsupervised learning* is to group $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ into clusters. *Data mining* is looking for relationships in large data sets.

3) For classical regression and multivariate analysis, we often want $n \geq 10p$, and a model with $n < 5p$ is overfitting: the model does not have enough data to estimate parameters accurately if $\boldsymbol{x}$ is $p \times 1$. Statistical Learning methods often use a model with a crude degrees of freedom d, where $n \geq Jd$ with $J \geq 5$ and preferably $J \geq 10$. A model is underfitting if it omits important predictors. Fix p, if the probability that a model underfits goes to 0 as the sample size $n \to \infty$, then overfitting may not be too serious if $n \geq Jd$. Underfitting can cause the model to fail to hold.

4) There are several important Statistical Learning principles.
i) There is more interest in prediction or classification, e.g. producing $\hat{Y}_f$, than in other types of inference.
ii) Often the focus is on extracting useful information when n/p is not large, e.g. $p > n$. If d is a crude estimator of the fitted model degrees of freedom, we want n/d large. A *sparse model* has few nonzero coefficients. We can have sparse population models and sparse fitted models. Sometimes sparse fitted models are useful even if the population model is *dense* (not sparse). Often the number of nonzero coefficients of a *sparse fitted model* $= d$.
iii) Interest is in how well the method performs on test data. Performance on training data is overly optimistic for estimating performance on test data.
iv) Some methods are *flexible* while others are *unflexible*. For unflexible methods, the sufficient predictor is often a hyperplane $SP = \boldsymbol{x}^T\boldsymbol{\beta}$ and often the mean function $E(Y|\boldsymbol{x}) = M(\boldsymbol{x}^T\boldsymbol{\beta})$ where the function M is known but the $p \times 1$ vector of parameters $\boldsymbol{\beta}$ is unknown and must be estimated (GLMs). Flexible methods tend to be useful for more complicated regression methods where $E(Y|\boldsymbol{x}) = m(\boldsymbol{x})$ for an unknown function m or $SP \neq \boldsymbol{x}^T\boldsymbol{\beta}$ (GAMs).

5) *Regression* investigates how the response variable Y changes with the value of a $p \times 1$ vector $\boldsymbol{x}$ of predictors. Often this *conditional distribution* $Y|\boldsymbol{x}$ is described by a *1D regression model*, where Y is conditionally independent of $\boldsymbol{x}$ given the *sufficient predictor* $SP = h(\boldsymbol{x})$, written

$$Y \perp\!\!\!\perp \boldsymbol{x}|SP \quad \text{or} \quad \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x}|\mathrm{h}(\boldsymbol{x}),$$

where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The *estimated sufficient predictor* ESP $= \hat{h}(\boldsymbol{x})$. An important special case is a model with a linear predictor $h(\boldsymbol{x}) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$

where ESP $= \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$. This class of models includes the *generalized linear model* (GLM). Another important special case is a *generalized additive model* (GAM), where Y is independent of $\boldsymbol{x} = (x_1, ..., x_p)^T$ given the *additive predictor* $AP = \alpha + \sum_{j=1}^p S_j(x_j)$ for some (usually unknown) functions S_j. The *estimated additive predictor* EAP = ESP = $\hat{\alpha} + \sum_{j=1}^p \hat{S}_j(x_j)$. The **response variable** is the variable that you want to predict. The **predictor** variables (or features) are used to predict the response variable.

6) Given a model know how to find the SP $h(\boldsymbol{x})$. Tip: if the model depends on $\boldsymbol{x}$ only through the real valued function $h(\boldsymbol{x})$, then $SP = h(\boldsymbol{x})$.

7) The **additive error regression** model is $Y = m(\boldsymbol{x}) + e$, suppressing subscripts. The model could be written $Y_i = m(\boldsymbol{x}_i) + e_i$ for $i = 1, ..., n$. The multiple linear regression (**MLR**) model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$ is a special case. The MLR model is

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$$

for $i = 1, \ldots, n$. Here n is the *sample size* and the random variable e_i is the ith **error**.

8) A *response plot* is a plot of ESP vs Y and a *residual plot* is a plot of ESP vs. r. For the models in 7), the ith residual $r_i = Y_i - \hat{m}(\boldsymbol{x})$, and the $ESP = \hat{m}(\boldsymbol{x}) = \hat{Y}$. If the errors are unimodal without much skew, then for models in 7) the plotted points should cluster about the identity line with unit slope and 0 intercept and the $r = 0$ line in the response and residual plots.

9) A plot of w vs. z puts w on the horizontal axis and z on the vertical axis.

10) A transformation model is $Y = t(Z) = m(\boldsymbol{x}) + e$. Assume that **all** of the values of the "response" Z_i are **positive**. A *power transformation* has the form $Y = t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $Y = t_0(Z) = \log(Z)$ for $\lambda = 0$ where

$$\lambda \in \Lambda_L = \{-1, -1/2, -1/3, 0, 1/3, 1/2, 1\},$$

the ladder of powers. A graphical method for response transformations computes the "fitted values" $\hat{W}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_\lambda$ from the multiple linear regression model using $W_i = t_\lambda(Z_i)$ as the "response." A *transformation plot* is a plot of $\hat{W}$ versus W with the identity line added as a visual aid. and is made for each of the seven values of $\lambda \in \Lambda_L$. The plotted points follow the identity line in a (roughly) evenly populated band if the iid error MLR model is reasonable for $Y = W$ and $\boldsymbol{x}$. Often TZHAT or YHAT is on the horizontal axis and $Y = t(Z)$ on the vertical axis.

11) Given several transformation plots or several response plots (with $Y = t(Z)$ or $t(Z)$ on the vertical axis), be able to find the response transformation $Y = t(Z)$ corresponding to a plot that looks like a good MLR response plot. Q1, HW1 C.

12) Suppose you have a scatterplot of two variables $x_1^{\lambda_1}$ versus $x_2^{\lambda_2}$, $x_1, x_2 > 0$ and that the plotted points follow a nonlinear one to one function. If $\lambda = 0$ use the log transformation $\log(x_i)$. Consider the **ladder of powers**. **Ladder rule:** To spread small values of the variable, make λ_i smaller. To spread large values of the variable, make λ_i larger. Be able to use the Ladder Rule.

13) Suppose that all values of the variable w to be transformed are positive. The **log rule** says use $\log(w)$ if $\max(w_i)/\min(w_i) > 10$. Be able to use the log rule.

14) Consider the ladder of powers given in point 10). No transformation ($\lambda = 1$) is best, then the log transformation, then the square root transformation.

15) Given a plot of x versus Y, be able to use the ladder rule to decide between two transformations, one decreasing λ, eg $\log(Y)$, and one increasing λ, eg Y^2. A variant might have a plot of $\sqrt{x}$ versus $\sqrt{Y}$. Then choose between Y and $\log(Y)$ or between x and $\log(x)$.

16) The *modified power transformation family*

$$t_\lambda(Z_i) \equiv Z_i^{(\lambda)} = \frac{Z_i^\lambda - 1}{\lambda} \tag{1}$$

for $\lambda \neq 0$ and $Z_i^{(0)} = \log(Z_i)$ where $\lambda \in \Lambda_L$.

17) For the location model, the sample mean $\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n}$, the sample variance $S_n^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1}$, and the sample standard deviation $S_n = \sqrt{S_n^2}$. If the data $Y_1, ..., Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the ith order statistic and the $Y_{(i)}$'s are called the *order statistics*. The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \ \text{ if n is odd,}$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{((n/2)+1)}}{2} \ \text{ if n is even.}$$

The notation $\text{MED}(n) = \text{MED}(Y_1, ..., Y_n)$ will also be used. The *sample median absolute deviation* is $\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|,\ i = 1, \ldots, n)$.

18) Suppose the multivariate data has been collected into an $n \times p$ matrix

$$\boldsymbol{W} = \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}.$$

The *coordinatewise median* $\text{MED}(\boldsymbol{W}) = (\text{MED}(X_1), ..., \text{MED}(X_p))^T$ where $\text{MED}(X_i)$ is the sample median of the data in column i corresponding to variable X_i. The **sample mean** $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i = (\overline{X}_1, ..., \overline{X}_p)^T$ where $\overline{X}_i$ is the sample mean of the data in column i corresponding to variable X_i. The **sample covariance matrix**

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T = (S_{ij}).$$

That is, the ij entry of $\boldsymbol{S}$ is the sample covariance S_{ij}. The *classical estimator of multivariate location and dispersion* is $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$.

19) Let $(T, \boldsymbol{C}) = (T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W}))$ be an estimator of multivariate location and dispersion. The ith *Mahalanobis distance* $D_i = \sqrt{D_i^2}$ where the ith *squared Mahalanobis distance* is $D_i^2 = D_i^2(T(\boldsymbol{W}), \boldsymbol{C}(\boldsymbol{W})) = (\boldsymbol{x}_i - T(\boldsymbol{W}))^T \boldsymbol{C}^{-1}(\boldsymbol{W})(\boldsymbol{x}_i - T(\boldsymbol{W}))$.

20) The squared Euclidean distances of the $\boldsymbol{x}_i$ from the coordinatewise median is $D_i^2 = D_i^2(\text{MED}(\boldsymbol{W}), \boldsymbol{I}_p)$. Concentration type steps compute the weighted median MED_j: the

coordinatewise median computed from the cases $\boldsymbol{x}_i$ with $D_i^2 \leq \text{MED}(D_i^2(\text{MED}_{j-1}, \boldsymbol{I}_p))$ where $\text{MED}_0 = \text{MED}(\boldsymbol{W})$. Often used $j = 0$ (no concentration type steps) or $j = 9$. Let $D_i = D_i(\text{MED}_j, \boldsymbol{I}_p)$. Let $W_i = 1$ if $D_i \leq \text{MED}(D_1, ..., D_n) + k\text{MAD}(D_1, ..., D_n)$ where $k \geq 0$ and $k = 5$ is the default choice. Let $W_i = 0$, otherwise.

21) Let the *covmb2 set* B of at least $n/2$ cases correspond to the cases with weight $W_i = 1$. Then the *covmb2* estimator $(T, \boldsymbol{C})$ is the sample mean and sample covariance matrix applied to the cases in set B. Hence

$$T = \frac{\sum_{i=1}^n W_i \boldsymbol{x}_i}{\sum_{i=1}^n W_i} \quad \text{and} \quad \boldsymbol{C} = \frac{\sum_{i=1}^n W_i(\boldsymbol{x}_i - T)(\boldsymbol{x}_i - T)^T}{\sum_{i=1}^n W_i \ - \ 1}.$$

The function `ddplot5` plots the Euclidean distances from the coordinatewise median versus the Euclidean distances from the `covmb2` location estimator. Typically the plotted points in this DD plot cluster about the identity line, and outliers appear in the upper right corner of the plot with a gap between the bulk of the data and the outliers.

22) If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $p \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}), \ E(\boldsymbol{a} + \boldsymbol{Y}) = \boldsymbol{a} + E(\boldsymbol{Y}), \ \& \ E(\boldsymbol{AXB}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}.$$

Also

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{AX}) = \text{Cov}(\boldsymbol{AX}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T.$$

Note that $E(\boldsymbol{AY}) = \boldsymbol{A}E(\boldsymbol{Y})$ and $\text{Cov}(\boldsymbol{AY}) = \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T$.

23) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

24) If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{AX} \sim N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants, then $\boldsymbol{X} + \boldsymbol{a} \sim N_p(\boldsymbol{\mu} + \boldsymbol{a}, \boldsymbol{\Sigma})$.

25) Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ as $n \to \infty$ for all points $\boldsymbol{x}$ at which $F(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \to 0$ as $n \to \infty$.

26) Multivariate Central Limit Theorem (MCLT): If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i.$$

27) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A\mu}) \xrightarrow{D} N_q(\boldsymbol{A\theta}, \boldsymbol{A\Sigma A}^T)$.

28) Suppose $\boldsymbol{A}$ is a conformable constant matrix and $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$. Then $\boldsymbol{AX}_n \xrightarrow{D} \boldsymbol{AX}$.

29) For $h > 0$, the hyperellipsoid $\{\boldsymbol{z} : (\boldsymbol{z} - T)^T \boldsymbol{C}^{-1} (\boldsymbol{z} - T) \leq h^2\} = \{\boldsymbol{z} : D^2_{\boldsymbol{z}} \leq h^2\} = \{\boldsymbol{z} : D_{\boldsymbol{z}} \leq h\}$. A future observation (random vector) $\boldsymbol{x}_f$ is in this region if $D_{\boldsymbol{x}_f} \leq h$. A large sample $100(1-\delta)\%$ prediction region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{x}_f \in \mathcal{A}_n) \xrightarrow{P} 1 - \delta$ where $0 < \delta < 1$. A large sample $100(1-\delta)\%$ confidence region is a set $\mathcal{A}_n$ such that $P(\boldsymbol{\mu} \in \mathcal{A}_n) \xrightarrow{P} 1 - \delta$. A prediction interval (PI) $[L_n, U_n]$ is a special case of a prediction region and a confidence interval (CI) $[L_n, U_n]$ is a special case of a confidence region.

30) Consider intervals that contain c cases $[Y_{(1)}, Y_{(c)}], [Y_{(2)}, Y_{(c+1)}], ..., [Y_{(n-c+1)}, Y_{(n)}]$. Compute $Y_{(c)} - Y_{(1)}, Y_{(c+1)} - Y_{(2)}, ..., Y_{(n)} - Y_{(n-c+1)}$. Then the estimator shorth(c) $= [Y_{(s)}, Y_{(s+c-1)}]$ is the interval with the shortest length. A large sample $100(1-\delta)\%$ prediction interval (PI) (L_n, U_n) is such that $P(Y_f \in (L_n, U_n)) \to 1 - \delta$ as $n \to \infty$. The shorth(c) interval is a large sample $100(1-\delta)\%$ PI if $c/n \to 1 - \delta$ as $n \to \infty$ that often has the asymptotically shortest length.

31) Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + p/n)$ for $\delta > 0.1$ and $q_n = \min(1 - \delta/2, 1 - \delta + 10\delta p/n)$, otherwise. If $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, d\boldsymbol{\Sigma})$, then $\{\boldsymbol{z} : D_{\boldsymbol{z}}(T, \boldsymbol{C}) \leq h\}$ is a large sample $100(1-\delta)\%$ prediction regions if $h = D_{(U_n)}$ where $D_{(U_n)}$ is the $100 q_n$th sample quantile of the D_i. The nonparametric prediction region uses $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \boldsymbol{S})$. See 35).

32) Suppose m independent large sample $100(1-\delta)\%$ prediction regions are made where $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ are iid from the same distribution for each of the m runs. Let Y count the number of times $\boldsymbol{x}_f$ is in the prediction region. Then $Y \sim$ binomial $(m, 1 - \delta_n)$ where $1 - \delta_n$ is the true coverage and $1 - \delta_n \to 1 - \delta$ as $n \to \infty$. Simulation can be used to see if the true or actual coverage $1 - \delta_n$ is close to the nominal coverage $1 - \delta$. A prediction region with $1 - \delta_n < 1 - \delta$ is liberal and a region with $1 - \delta_n > 1 - \delta$ is conservative. It is better to be conservative by 5% than liberal by 5%. Parametric prediction regions tend to have large undercoverage and so are too liberal.

33) For the nonparametric prediction region, we want $n \geq 10p$ for good coverage and $n \geq 50p$ for good volume.

34) Let q_n and c be given by 31) with p replaced by d, a crude estimator of the model degrees of freedom. Let

$$b_n = \left(1 + \frac{15}{n}\right) \sqrt{\frac{n + 2d}{n - d}} \quad \text{if} \;\; \text{d} \leq \text{8n/9}, \quad \text{and} \quad \text{b}_\text{n} = 5\left(1 + \frac{15}{\text{n}}\right),$$

otherwise. Compute the shorth(c) of the residuals $= [r_{(s)}, r_{(s+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$. Let $\hat{Y}_f = \hat{m}(\boldsymbol{x}_f)$. Then a 100 $(1-\delta)\%$ large sample PI for Y_f is

$$[\hat{m}(\boldsymbol{x}_f) + b_n \tilde{\xi}_{\delta_1}, \hat{m}(\boldsymbol{x}_f) + b_n \tilde{\xi}_{1-\delta_2}].$$

Note that this PI roughly uses the shorth of the pseudodata $\hat{Y}_f + r_i$ for $i = 1, ..., n$.

35) Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{c}$ where $\boldsymbol{c}$ is a known $r \times 1$ vector. The **prediction region method** makes a bootstrap sample $\boldsymbol{w}_i = \hat{\boldsymbol{\mu}}^*_i - \boldsymbol{c}$ for $i = 1, ..., B$. Make the nonparametric prediction region $\{\boldsymbol{z} : D^2_{\boldsymbol{z}}(\overline{\boldsymbol{x}}, \boldsymbol{S}) \leq D^2_{(U_n)}\}$ for the $\boldsymbol{w}_i$, and reject H_0 if $\boldsymbol{0}$ is not in the prediction region. See 31).