# STAT-6655 HWK 1

Cody Grogan
A02313514

January 22, 2024

## 1) Two Cultures

The main points in this paper are related to the two classes of statistical modeling, data modeling and algorithmic modeling. For data modeling, the author emphasizes quality of fit doesn't mean quality of prediction, there are often many fits that produce the same error, and making assumptions about a problem often limits the answer. For the first point, the author talks at great length about the misuse of the "5% significance" test used in early data modeling. This test was only meant for extremely low variable fits and often lead researchers to derive conclusions about nature from models that could not predict future events well. The next point was the multiplicity problem for data models, more importantly, in a given dataset, there are often multiple data models that can produce the same measure of error. The author then shows this in an example where 2 high importance variables in a classification problem produced comparable accuracy with either removed from the regression. Finally the author touches on how assumptions of normally distributed or IID data often limits the quality of a model to the point where such models have give questionable results.

For algorithmic modeling, the main points were simple is not always better, high dimensionality is not always bad, and a model that can predict with high accuracy holds more information about the underlying model. On the point that simple is not always better, the author emphasises processes in nature are often not simple so a simplified model will give questionable conclusions. I agree with this point because if a process in nature is simple enough to model with regression then there would be no need for regression at all. Or, at the very least, regression models would have high validation accuracy. The next point the author writes about for algorithmic modeling is dimensionality is not always a curse. More specifically, with his SVM example, the author shows increasing the dimensionality of a problem can yield higher accuracy. The last point about algorithmic modeling the author emphasises is that a model with better prediction accuracy is one that models the underlying process best. I agree with this point as well prediction on future points is a great indicator of the knowledge of a model even if that model is less interpretable. Moreover, it is better to create these high accuracy models and find a way to extract information about the model then to use a highly interpretable model.

## 2) Probability

### a)

#### i

Using integral notation for the expected value of random variables,

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyp(x,y)dxdy \tag{1}$$

Then using the assumption of independence,

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyp(x)p(y)dxdy \tag{2}$$

Now because of the independence, the double integral can be evaluated as the product of 2 single integrals,

$$E[XY] = \int_{-\infty}^{\infty} xp(x)dx \int_{-\infty}^{\infty} yp(y)dy \tag{3}$$
$$E[XY] = E[X]E[Y] \tag{4}$$

**ii**

Starting with integral notation again,

$$E[X+Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)p(x,y)dxdy \tag{5}$$

Distributing the pdf and using the additive principle of integrals,

$$E[X+Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xp(x,y) + yp(x,y)dxdy \tag{6}$$
$$E[X+Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xp(x,y)dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yp(x,y)dxdy \tag{7}$$
$$E[X+Y] = E[X] + E[Y] \tag{8}$$

**iii**

Defining the variance and expanding,

$$Var[X+Y] = E[((X+Y) - E[X+Y])^2] \tag{9}$$
$$Var[X+Y] = E[(X+Y)^2 - 2(X+Y)E[X+Y] + E[X+Y]^2] \tag{10}$$

Now simplifying using the previous summation derivation and an expectation is a constant,

$$Var[X+Y] = E[(X+Y)^2] - 2E[X+Y]E[X+Y] + E[X+Y]^2 \tag{11}$$
$$Var[X+Y] = E[(X+Y)^2] - E[X+Y]^2 \tag{12}$$

Now expanding inside the first expectation,

$$Var[X+Y] = E[X^2 + 2XY + Y^2] - E[X+Y]^2 \tag{13}$$

Now distributing the expectation and using independence

$$Var[X+Y] = E[X^2] + 2E[X]E[Y] + E[Y^2] - E[X+Y]^2 \tag{14}$$

Now using the summation rule of expectations from ii and expanding the second term

$$Var[X+Y] = E[X^2] + 2E[X]E[Y] + E[Y^2] - (E[X] + E[Y])^2 \tag{15}$$
$$Var[X+Y] = E[X^2] + 2E[X]E[Y] + E[Y^2] - (E[X]^2 + 2E[X]E[Y] + E[Y]^2) \tag{16}$$
$$Var[X+Y] = E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 \tag{17}$$
$$Var[X+Y] = Var[X] + Var[Y] \tag{18}$$

**iv**

Writing out the integral form of the right side,

$$E_Y[E_X[X|Y]] = \int_{-\infty}^{\infty} (\int_{-\infty}^{\infty} x p_{X|Y}(X|Y) dx) p_Y(y) dy \tag{19}$$

Now distributing the outside integral into the inside,

$$E_Y[E_X[X|Y]] = \iint_{-\infty}^{\infty} x p_Y(y) p_{X|Y}(X|Y) dx dy \tag{20}$$

Now using the definition of the conditional pdf,

$$E_Y[E_X[X|Y]] = \iint_{-\infty}^{\infty} x p_Y(y) \frac{p(x,y)}{p_Y(y)} dx dy \tag{21}$$

$$E_Y[E_X[X|Y]] = \iint_{-\infty}^{\infty} x p(x,y) dx dy \tag{22}$$

Now switching the order of the integration,

$$E_Y[E_X[X|Y]] = \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} p(x,y) dy dx \tag{23}$$

$$E_Y[E_X[X|Y]] = \int_{-\infty}^{\infty} x p(x) dx \tag{24}$$

$$E_Y[E_X[X|Y]] = E[X] \tag{25}$$

**v**

Assuming the indicator function follows the form,

$$1(x) \begin{cases} 0 & x \notin C \\ 1 & x \in C \end{cases} \tag{26}$$

The expectation for the continuous random variable can then be written as,

$$E[1|X \in C] = \int_C 1(x) p(x) dx \tag{27}$$

Now the integral over the set C can be split into 2 arbitrary disjoint subsets where $A \cup B = C$ and $A \cap B = \emptyset$.

$$E[1|X \in C] = \int_A 1(x) p(x) dx + \int_B 1(x) p(x) dx \tag{28}$$

Then if we assume that A is the subset of C that does not contain X and B is the subset that contains X,

$$E[1|X \in C] = \int_A 0 \cdot p(x) dx + \int_B 1 \cdot p(x) dx \tag{29}$$

$$E[1|X \in C] = \int_B p(x) dx \tag{30}$$

$$E[1|X \in C] = P(X \in C) \tag{31}$$

### b)

$$p(x|u) = \left(\frac{1-u}{2}\right)^{(1-x)/2} \left(\frac{1+u}{2}\right)^{(1+x)/2} \tag{32}$$

Following the rules of a valid pmf we know $p(x|u) \geq 0 \; \forall x$ and $\sum_x p(x|u) = 1$. These translate into the following inequalities

$$\frac{1-u}{2} \geq 0 \tag{33}$$
$$\frac{1+u}{2} \geq 0 \tag{34}$$
$$\frac{1-u}{2} + \frac{1+u}{2} = 1 \tag{35}$$

Now simplifying these,

$$1 \geq u \geq -1 \tag{36}$$
$$2 = 2 \tag{37}$$

This then shows the valid interval for u is [-1,1]. Finding the expected value of the pmf,

$$E[X] = \sum_x xp(x|u) \tag{38}$$
$$E[X] = (-1)p(X = -1|u) + (1)p(X = 1|u) \tag{39}$$
$$E[X] = -\left(\frac{1-u}{2}\right) + \left(\frac{1+u}{2}\right) \tag{40}$$
$$E[X] = \frac{u-1}{2} + \frac{1+u}{2} \tag{41}$$
$$E[X] = u \tag{42}$$

Finding the variance of the pmf,

$$Var[X] = E[X^2] - E[X]^2 \tag{43}$$
$$Var[X] = \sum_x x^2 p(x|u) - u^2 \tag{44}$$
$$Var[X] = (-1)^2 \left(\frac{1-u}{2}\right) + (1)^2 \left(\frac{1+u}{2}\right) - u^2 \tag{45}$$
$$Var[X] = \frac{1-u}{2} + \frac{1+u}{2} - u^2 \tag{46}$$
$$Var[X] = 1 - u^2 \tag{47}$$

## 3) Linear Algebra

### a)

Using the identity,

$$U\Lambda U^T = \sum_{i=1}^{d} \lambda_i u_i u_i^T \tag{48}$$

$$x^T Ax = x^T (\sum_{i=1}^{d} \lambda_i u_i u_i^T) x \tag{49}$$

$$x^T Ax = \sum_{i=1}^{d} \lambda_i x^T u_i u_i^T x \tag{50}$$

Now we know that $x^T u_i$ is simply the dot product of the 2 vectors and is equal to a scalar we will call v,

$$x^T Ax = \sum_{i=1}^{d} \lambda_i x^T u_i (x^T u_i)^T \tag{51}$$

$$x^T Ax = \sum_{i=1}^{d} \lambda_i v (v)^T \tag{52}$$

$$x^T Ax = \sum_{i=1}^{d} \lambda_i v^2 \tag{53}$$

We then know that $v^2$ is always a positive real number so for the matrix to be PSD $\lambda \geq 0 \; \forall \lambda \in A$.

## b)

Starting with the spectral decomposition,

$$A = U\Lambda U^T \tag{54}$$

Now since U is an orthogonal matrix we know that its transpose is equal to its inverse,

$$AU = U\Lambda \tag{55}$$

Now writing the columns of the resulting matrices,

$$\begin{bmatrix} Au_1 & Au_2 & \dots & Au_n \end{bmatrix} = U\Lambda \tag{56}$$

Now since $\Lambda$ is a diagonal matrix the matrix multiplication equates to the ith eigenvector times the ith eigenvalue,

$$\begin{bmatrix} Au_1 & Au_2 & \dots & Au_n \end{bmatrix} = \begin{bmatrix} \lambda_1 u_1 & \lambda_2 u_2 & \dots & \lambda_n u_n \end{bmatrix} \tag{57}$$

We can see each column is the definition for an eigenvector $Ax = \lambda x$. Thus $u_i$ is the ith eigenvector of A with and eigenvalue of $\lambda_i$

## c)

The definition of a PD matrix is,

$$x^T Ax > 0 \tag{58}$$

From this we know,

$$\exists a > 0 \mid x^T A x = a \ \forall x \tag{59}$$
$$\exists b > 0 \mid x^T A x = b \ \forall x \tag{60}$$

Adding these equations together we get,

$$x^T A x + x^T B x = a + b \tag{61}$$
$$x^T (A + B) x = a + b \tag{62}$$

Thus we know that the sum of two PD matrices is also PD because,

$$a > 0 \ and \ b > 0 \implies a + b > 0 \tag{63}$$

### f)

Squaring both sides of the equation,

$$||x||^2 = ||Ux||^2 \tag{64}$$
$$x^T x = (Ux)^T (Ux) \tag{65}$$
$$x^T x = x^T U^T U x \tag{66}$$
$$x^T x = x^T x \tag{67}$$

## 4) Multivariate Calculus

Starting with the equation,

$$l(u, \Sigma, x_1, ..., x_n) = \log \left( \prod_i^n \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} exp \left[ -\frac{1}{2} (x_i - u)^T \Sigma^{-1} (x_i - u) \right] \right) \tag{68}$$

Now using the rules of logarithms,

$$l(u, \Sigma, x_1, ..., x_n) = \sum_i^n \log \left( \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} exp \left[ -\frac{1}{2} (x_i - u)^T \Sigma^{-1} (x_i - u) \right] \right) \tag{69}$$

Using the chain rule for derivatives and defining the following terms

$$v = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} exp \left[ -\frac{1}{2} (x_i - u)^T \Sigma^{-1} (x_i - u) \right] \tag{70}$$
$$w = -\frac{1}{2} (x_i - u)^T \Sigma^{-1} (x_i - u) \tag{71}$$
$$y = x_i - u \tag{72}$$

Now defining the chain rule,

$$\frac{\partial l}{\partial u} = \sum_i^n \frac{\partial l}{\partial v} \frac{\partial v}{\partial w} \frac{\partial w}{\partial y} \frac{\partial y}{\partial u} \tag{73}$$
$$\frac{\partial l}{\partial u} = \sum_i^n \nabla_u (x_i - u) \nabla_y (-\frac{1}{2} y^T \Sigma^{-1} y) \frac{\partial}{\partial w} (\frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} exp\,[w]) \frac{\partial}{\partial v} (\log(v)) \tag{74}$$
$$\frac{\partial l}{\partial u} = \sum_i^n (-I)(-\Sigma^{-1} y)(\frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} exp\,[w])(\frac{1}{v}) \tag{75}$$

Now simplifying because $w$ is a scalar value,

$$\frac{\partial l}{\partial u} = \sum_i^n \Sigma^{-1}(x_i - u)(\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}(x_i - u)^T\Sigma^{-1}(x_i - u)\right])((2\pi)^{p/2}|\Sigma|^{1/2} exp - \left[-\frac{1}{2}(x_i - u)^T\Sigma^{-1}(x_i - u)\right]) \tag{76}$$

$$\frac{\partial l}{\partial u} = \sum_i^n \Sigma^{-1}(x_i - u) \tag{77}$$

Setting this equal to 0 and solving for u,

$$0 = \sum_i^n \Sigma^{-1}(x_i - u) \tag{78}$$

$$0 = \sum_i^n \Sigma^{-1}x_i - \sum_i^n \Sigma^{-1}u \tag{79}$$

$$\sum_i^n \Sigma^{-1}u = \sum_i^n \Sigma^{-1}x_i \tag{80}$$

$$n\Sigma^{-1}u = \Sigma^{-1}\sum_i^n x_i \tag{81}$$

$$u = \frac{1}{n}\sum_i^n x_i \tag{82}$$

# 5) Estimation

## a)

Defining the bias for the estimator,

$$B = E[\bar{X}] - \mu \tag{83}$$

$$B = E[\frac{1}{n}\sum_i^n X_i] - \mu \tag{84}$$

$$B = \frac{1}{n}\sum_i^n E[X_i] - \mu \tag{85}$$

$$B = \frac{1}{n}\sum_i^n \mu - \mu \tag{86}$$

$$B = \mu - \mu = 0 \tag{87}$$

## b)

Using the equation for variance,

$$Var[\bar{X}] = Var[\frac{1}{n}\sum_{i}^{n} X_i] \tag{88}$$

$$Var[\bar{X}] = \frac{1}{n^2} Var[\sum_{i}^{n} X_i] \tag{89}$$

$$Var[\bar{X}] = \frac{1}{n^2}\sum_{i}^{n} Var[X_i] \tag{90}$$

$$Var[\bar{X}] = \frac{1}{n^2}\sum_{i}^{n} \sigma^2 \tag{91}$$

$$Var[\bar{X}] = \frac{\sigma^2}{n} \tag{92}$$

## c)

Defining the MSE of the estimator,

$$MSE[\bar{X}] = E[(\bar{X} - \mu)^2] \tag{93}$$

Now expanding this I get,

$$MSE[\bar{X}] = E[\bar{X}^2 - 2\bar{X}\mu + \mu^2] \tag{94}$$

$$MSE[\bar{X}] = E[\bar{X}^2] - 2E[\bar{X}]\mu + \mu^2 \tag{95}$$

$$MSE[\bar{X}] = E[\bar{X}^2] - \mu^2 \tag{96}$$

Now using the definition for variance,

$$Var[\bar{X}] = E[\bar{X}^2] - E[\bar{X}]^2 \tag{97}$$

$$E[\bar{X}^2] = Var[\bar{X}] + E[\bar{X}]^2 \tag{98}$$

Substituting this into the MSE equation,

$$MSE[\bar{X}] = Var[\bar{X}] + E[\bar{X}]^2 - \mu^2 \tag{99}$$

$$MSE[\bar{X}] = \frac{\sigma^2}{n} + \mu^2 - \mu^2 \tag{100}$$

$$MSE[\bar{X}] = \frac{\sigma^2}{n} \tag{101}$$

## d)

Defining the bias of the estimator,

$$B = E[\hat{s}^2] - \sigma^2 \tag{102}$$

$$B = E[\frac{1}{n}\sum_{i}^{n}(X_i - \bar{X})^2] - \sigma^2 \tag{103}$$

$$B = \frac{1}{n}\sum_{i}^{n} E[X_i^2 - 2X_i\bar{X} + \bar{X}^2] - \sigma^2 \tag{104}$$

$$B = \frac{1}{n}\sum_{i}^{n} \left(E[X_i^2] - 2E[X_i\bar{X}] + E[\bar{X}^2]\right) - \sigma^2 \tag{105}$$

Distributing the sum,

$$B = \frac{1}{n}\left(\sum_i^n E[X_i^2] - 2\sum_i^n E[X_i\bar{X}] + \sum_i^n E[\bar{X}^2]\right) - \sigma^2 \tag{106}$$

$$B = \frac{1}{n}\left(\sum_i^n E[X_i^2] - 2E[\bar{X}\sum_i^n X_i] + \sum_i^n E[\bar{X}^2]\right) - \sigma^2 \tag{107}$$

$$B = \frac{1}{n}\left(\sum_i^n E[X_i^2] - 2nE[\bar{X}^2] + nE[\bar{X}^2]\right) - \sigma^2 \tag{108}$$

$$B = \frac{1}{n}\left(\sum_i^n E[X_i^2] - nE[\bar{X}^2]\right) - \sigma^2 \tag{109}$$

$$B = \frac{1}{n}\left(\sum_i^n E[X_i^2] - nE[\bar{X}^2]\right) - \sigma^2 \tag{110}$$

Now using the equation for variance again,

$$B = \frac{1}{n}\left(n(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2)\right) - \sigma^2 \tag{111}$$

$$B = \frac{1}{n}\left(n\frac{n\sigma^2 - \sigma^2}{n}\right) - \sigma^2 \tag{112}$$

$$B = \frac{1}{n}\left(n\frac{\sigma^2(n-1)}{n}\right) - \sigma^2 \tag{113}$$

$$B = \frac{\sigma^2(n-1)}{n} - \sigma^2 \tag{114}$$

$$B = \frac{-\sigma^2}{n} \tag{115}$$

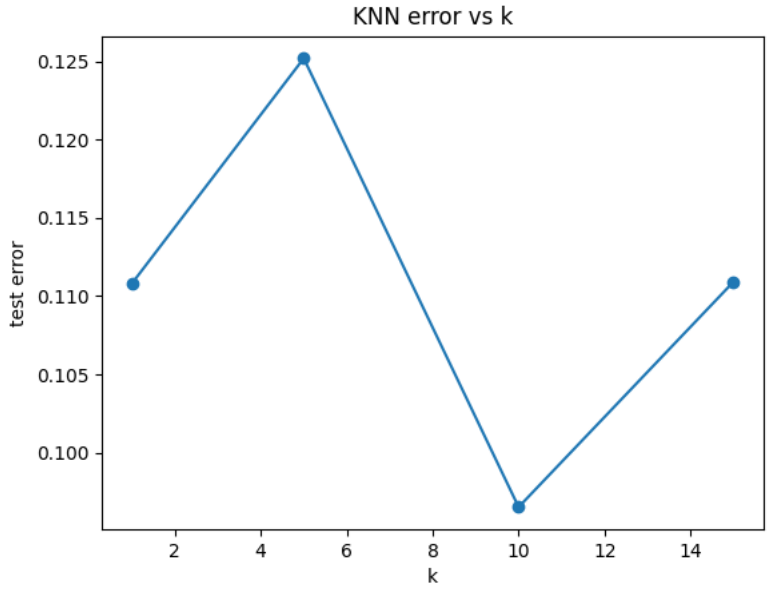
To make an unbiased estimator one could change the coefficient to $\frac{1}{n-1}$. Looking at equation 113, this would result in

$$B = \frac{1}{n-1}\left(n\frac{\sigma^2(n-1)}{n}\right) - \sigma^2 \tag{116}$$

$$B = \sigma^2 - \sigma^2 = 0 \tag{117}$$

# 6) KNN Classifier

The following is a plot of the testing error vs the number of neighbors of the Sci-kit KNN Classifier with 5-fold cross validation,




From the figure, the KNN classifier with 10 neighbors performs the best. A more detailed sweep of the neighbor parameter reveals the smallest testing error is achieve with 9 or 10 neighbors.