Data Science (CS 148)
February 15, 2023

Midterm Exam

Duration: 90 minutes

Write your name and UID:

Note 1: If you find a question difficult, move on with the rest of the questions and come back to it in the end!

Note 2: If you need more space for a question, use back of page 5.

1 Linear Regression (20 points)

Suppose researchers develop a drug to treat anxiety. We have measured the anxiety score vs. dose of the drug, from a sample of males and females in the population who are using the drug. The following plot shows the mean anxiety score by gender for each dose and connect the means with lines.

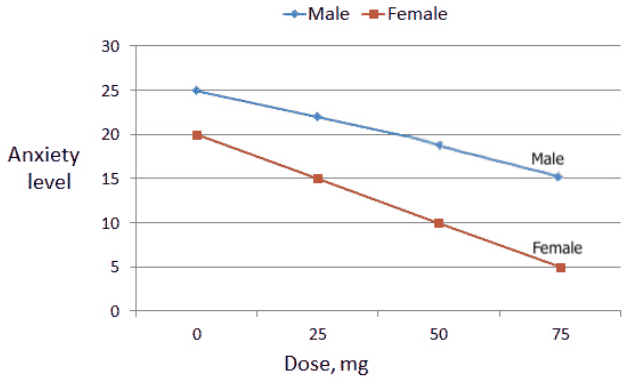


Figure 1: Mean anxiety level (y-axis) vs. dose (x-axis) for males (top line) and females (bottom line).

(a) **(5 points)** How do you model mean anxiety based on gender *and* dose, using *one* linear regression model? Write the formulation of your linear regression model in 1 line. *Note: You do not need to calculate the value of the coefficients.*

Solution: anxiety $= \beta_0 + \beta_1 \text{dose} + \beta_2 \text{gender} + \beta_3 \text{dose} \times \text{gender}$

(b) **(8 points)** Write the interpretation of each of the coefficients in your model.

Solution:

For female, gender $= 1$

For males: anxiety $= \beta_0 + \beta_1 \text{dose}$

For females: anxiety $= (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\text{dose}$

β_0: mean anxiety for dose=0 for males

β_1: for 1 mg increase in dose, anxiety decreases

β_2: difference between mean anxiety for dose $= 0$ between males and females

β_3: for 1 mg increase in dose, anxiety decreases by β_3 more for females

(c) **(7 points)** Calculate the coefficients of your model from the figure. *Note: you can leave fractions as is.*

Solution:

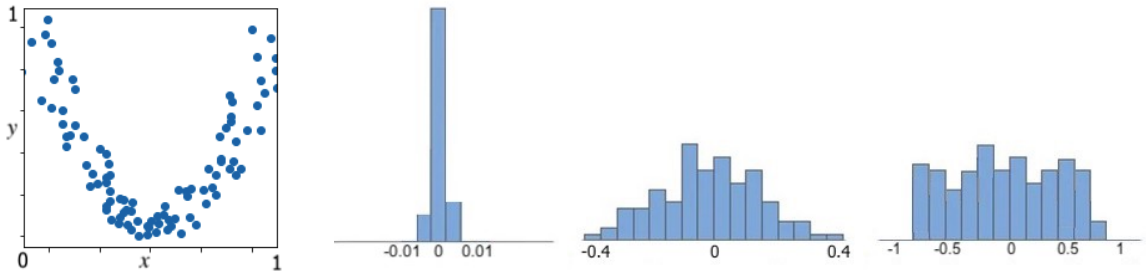
Two set of solutions are acceptable answers, based on how you determine gender $= 1$.

If for female we let gender $= 1$: $\beta_0 = 25, \beta_2 = -5, \beta_1 = -\frac{2}{15} = -0.13, \beta_3 = -0.2 - (-0.13) = -\frac{1}{15} = -0.07$;

If for male we let gender $= 1$: $\beta_0 = 20, \beta_2 = 5, \beta_1 = -\frac{1}{5}, \beta_3 = \frac{1}{15}$;

2 Model Selection & Bias-Variance Trade-off (16 points)

We use L_1-regularized linear regression with higher order terms to model the data in Fig. 2a. Formally, the model is: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k$ and we add a L_1-regularization term, i.e., $R = \sum_{i=1}^{k} |\beta_i|$, to the original MSE loss and find βs by minimizing $L_{reg}(\beta) = L_{MSE}(\beta) + \lambda R(\beta)$.



(a) Data (b) Model 1 (λ_1) (c) Model 2 (λ_2) (d) Model 3 (λ_3)

Figure 2: (a) Data, (b), (c), (d) Three models with different values of λ (coefficient of L_1 regularization).

(a) **(4 points)** Fig. 2b, 2c, 2d show the distribution of residuals for 3 different values of λ. Which models (Fig. 2b, or 2c, or 2d) correspond to largest and smallest λ respectively? Explain briefly.

Solution: Model 1: smallest λ i.e. regularized least. This model has nearly 0 residuals, so model is most complex.

Model 3: largest λ i.e. regularized most. This model has largest residuals so model is least complex.

Instead of model complexity, can reference underfitting/overfitting or bias/variance or MSE dominates / regularization dominates for full credit.

(b) **(4 points)** Which model (Fig. 2b, or 2c, or 2d) has the largest variance and which one has the largest bias? Explain briefly.

Solution: Model 1: largest variance since it is able to get nearly 0 residuals for all data points (overfitting). Model 3: largest bias since it is the least expressive model (underfitting).

(c) **(4 points)** Which model would you use and what is the issue with those that you don't chose?

Solution: Model 2, because it has the right complexity (small bias and variance). Model 1 overfits (because although the data is quadratic, it has noise and hence a model with a quadratic complexity cannot achieve 0 residuals for most points) and Model 3 underfits (because it has a similar number of large and small residuals), so both Model 1 and 3 don't have good generalization.

(d) **(4 points)** Explain the effect of L_1 regularization on the above model, and discuss briefly why L_1 it a good choice when modeling the data in Fig. 2a using several higher-order terms?

Solution: L_1 makes some of the coefficients exactly 0. The data is quadratic here, so we do not need the terms with a larger power than 2. L_1 will push the model towards zeroing out most coefficients and this will likely zero out the higher order terms. Thus, L_1 regularization helps train a model of the right complexity for this problem (References to preventing overfitting or large variance will also work here).

3 Logistic Regression & Decision Boundary (30)

We want to use Logistic regression to model the probability for an individual to be sick ($Y = 1$) based on whether the individual is vaccinated ($X_1 = 1$) or not vaccinated ($X_1 = 0$).

(a) **(4 points)** Write the logistic regression formulation to model log (use ln) odds of being sick, based on whether the individual is vaccinated or not.

Solution: $\ln(\frac{Y=1}{Y=0}) = \beta_0 + \beta_1 X_1$.

(b) **(4 points)** In our population, 20% of vaccinated individuals and 40% of unvaccinated individuals are sick. Calculate the intercept (β_0) and the coefficient for X_1 (β_1). *Note: write your answer based on* "ln".

Solution: For $X_1 = 0$, we have $\beta_0 = \ln\left(\frac{P(Y=1|X=0)}{P(Y=0|X=0)}\right) = \ln(.4/.6) = \ln(2/3)$. For $X_1 = 1$, we have $\beta_1 = \ln\left(\frac{P(Y=1|X=1)}{P(Y=0|X=1)}\right) - \beta_0 = \ln(.2/.8) - \ln(.4/.6) = \ln(1/4) - \ln(2/3) = \ln(3/8) = -0.98$.

Your solution don't have to be the exact numerical value. Just using ln with the correct fractional number will be fine. e.g. both $\ln(.4/.6)$ and $\ln(2/3)$ are acceptable for β_0.

(c) **(5 points)** Interpret β_0 and β_1. What is the *percentage* of change in odds of being sick as a result of vaccination? *Note: write your answer using* "ln".

Solution: β_0: log odds of being sick when a person is not vaccinated.

β_1: the difference in log odds of being sick for vaccinated and unvaccinated person. Or e^{β_1} is the odd ratio of vaccinated individuals getting sick over unvaccinated individuals. I.e., vaccinated individuals have $e^{\beta_1} = e^{-0.98} = 0.37$ times (63% less) odds of being sick.

The percentage is $(1 - e^{\beta_1}) \times 100$, with β_1 as computed in (b). As long as the formulation is correct, we won't penalize you here for incorrect answer in (b).

(d) **(5 points)** If the 95% confidence interval for β_1 is $[-2.8, 0.2]$, interpret the effect of vaccine on the probability of getting sick.

Solution: Getting the vaccine decreases the odds of being sick by 63%. However, this is not statistically significant (as the confidence interval contains 0). Hence, we are not confident that vaccination has an impact on the probability of getting sick.

(e) **(6 points)** We want to include another (real valued) variable X_2 in the model to also captures the effect of "the amount of exercise per week (in hours)" on the probability of getting sick. In your model, include X_2 with coefficient β_2, and an interaction term X_1X_2 with coefficient β_3. If $\beta_0 = 0.6, \beta_1 = -0.2, \beta_2 = -0.3, \beta_3 = -0.1$, write the formulation for the decision boundary. What is the minimum amount of exercise per week that a vaccinated and an unvaccinated individual need to do to not get sick?

Solution:

$$0.6 - 0.2X_1 - 0.3X_2 - 0.1X_1X_2 = 0$$

For $X_1 = 0$, we have $0.6 - 0.3X_2 = 0$ hence $X_2 = 2$. Above 2 hours, we predict 0. For $X_1 = 1$, we have $0.6 - 0.2 - 0.3X_2 - 0.1X_2 = 0$, hence $0.4 - 0.4X_2 = 0$ which implies $X_2 = 1$. Above 1 hour we predict 0. Therefore, the minimum amount of exercise for vaccinated person is 1 hour and for unvaccinated person is 2 hours.

(f) **(6 points)** Interpret β_3 by comparing its effect on odds of being sick for vaccinated vs unvaccinated group.

Solution: For unvaccinated individuals, the odd ratio is $e^{-0.3}$ for one hour per week increase in exercise. For vaccinated individuals, the odd ratio is $e^{-0.4}$ for one hour per week increase in exercise. The ratio of these two odd ratios (for vaccinated vs unvaccinated individuals) is the exponentiated coefficient for the interaction term, i.e. $e^{-0.1} = e^{-0.4}/e^{-0.3}$.

4 KNN (16 points)

Fig. 3 shows a two-dimensional data set with two classes (+) and (-). An additional point (?) on the graph is currently unlabeled.

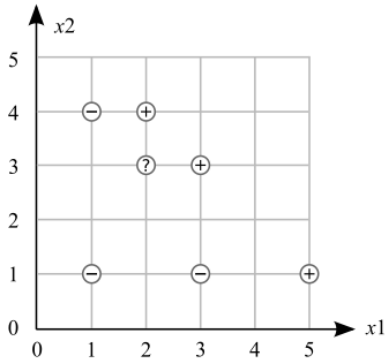


Figure 3: 2 Dimensional Array with labeled points

(a) **(6 points)** Using Manhattan distance (sum of absolute differences in each dimension of x_1 and x_2), determine the class of the unlabeled point for the following K values in the table below.

K Value	Classification?
1	**Solution: +**
3	**Solution: +**
5	**Solution: -**

(b) **(4 points)** If the scale of x_1 is very different than the scale of x_2, do you expect the model to perform well? Explain briefly. If your answer is no, what do you do to improve the model's performance?

Solution: No. KNN models are highly dependent on the relative scale of features in determining overall proximity. Failure to scale features will result in the model disproportionately weighing features of greater relative scale

(c) **(3 points)** If your data has many features $(x_1, x_2, \cdots, x_k)$, do you expect your KNN to work well? Explain briefly.

Solution: KNN works well with a small number of input features, but struggles when the number of inputs is very large (distances become similar)

(d) **(3 Points)** How do you expect the value of K in a KNN model to impact the variance of your model?

Solution: When you decrease the k the variance will typically increase

5 Classification Metrics (18 points)

We train a binary classifier to predict if a patient has cancer ($Y = 1$). We use the output probability of the classifier $P(Y = 1|X)$ to predict $Y = 1$ for a patient X, if $P(Y = 1|X) \geq 0.5$, and we predict $Y = 0$ otherwise. The classifier has following confusion matrix.

	Predicted $\hat{Y} = 1$	Predicted $\hat{Y} = 0$
Actual $Y = 1$	4	96
Actual $Y = 0$	1	9899

(a) **(8 points)** Calculate the **Accuracy, Precision, Recall** and **F1 Score** and of the classifier.

Solution: Accuracy=$\frac{9899+4}{9899+5+96} = 0.9903$; Precision=$\frac{4}{4+1} = 0.8$; Recall=$\frac{4}{96+4} = 0.04$; F1 Score: 0.076

(b) **(4 points)** Explain the discrepancy between accuracy and F1 score. Based on the above confusion matrix, which metric is better to evaluate the performance of the classifier?

Solution: Dataset is imbalanced. There are many examples with $Y = 0$ and only a few examples with $Y = 1$ (also many True Negatives but not too many True Positives). Thus, although the model is a bad predictor, the accuracy is still high. However, since the F1 score measures both False Positives and False Negatives, it can identify this model as a bad model with a low F1 score. Based on the above confusion matrix, the F1 score is a better metric.

(c) **(3 points)** How do you use the output probabilities of the above classifier to predict $Y = 1$ for a cancer screening test, where it is most important to identify *all* the possible cancer patients?

Solution: Using a lower threshold for predicting $Y = 1$, i.e., $P(Y = 1|X) \geq \pi$ for $\pi < 0.5$ will make sure that patients with a smaller chance of getting cancer will still be classified as having cancer, thus we can identify all cancer patients.

(d) **(3 points)** How do you use the output probabilities of the above classifier to predict $Y = 1$ for a follow-up test after treatment, where we don't want to tell a patient they're clear if this is not actually the case (i.e., we want to have fewer False Negatives)?

Solution: Using a lower threshold for predicting $Y = 1$, i.e., $P(Y = 1|X) \geq \pi$ for $\pi < 0.5$ (equivalently increasing the threshold for predicting $Y = 0$) will make sure that patients with a higher chance of not having cancer will still be classified as having cancer, thus making sure everyone that is clear more likely to be clear.