sample midterm

Instructions

Answer as many questions as you can. Be brief but include enough details to let me follow your reasoning.

- Notes, textbook, and calculator are allowed; computer is only allowed for Q2 and course materials.
- Each problem is 20 pts. Total points = 40. Time = 1 hr

1 Estimate a missing value.

Do the following by hand…

A sample of size $n = 6$ contains two independent variables X_1, X_2 and one categorical response variable Y. However, the response value for the 3rd sampling unit is missing.

	1	2	3	4	5	6
X_1	-2	-1	0	1	2	3
X_2	2	-2	0	0	3	2
Y	A	A	???	B	B	A

Predict the missing $Y_{i=3}$ by the following classification methods.

a. Predict $Y_{i=3}$ by KNN method with $k = 1$.

b. Predict $Y_{i=3}$ by KNN method with $k = 3$.

c. Logistic regression, without the 3rd sampling unit, produced the following results.

```
> z = 1*(y=="A")

> lreg = glm( z ~ x1 + x2, family=binomial)

> summary(lreg)

Coefficients: Estimate

Intercept 0.8373

x1 -0.4494

x2 -0.0778
```

- Estimate the probability $Y_{i=3} = A$.

d. Suppose X_1 and X_2 are independent Normal random variables with the same variance $\sigma^2 = 1$. Assuming equal prior probabilities $P(A) = P(B) = 0.5$, predict $Y_{i=3}$ using linear discriminant analysis.

e. Estimate the training error rate of the $k = 3$ KNN algorithm used in b.

f. (Stat-627 only) Estimate the testing error rate of the algorithm used in b using leave-one-out cross-validation.

2 Toothgrowth

An experiment was conducted to evaluate the effect of vitamin C on tooth growth. Sixty guinea pigs received various doses of vitamin C by one of two delivery methods, orange juice or ascorbic acid.

- Results of this experiment are in dataset `ToothGrowth` which is already loaded in R.
- You can look at it with commands `names(ToothGrowth)`, `summary(ToothGrowth)`, `dplyr::gimpse(ToothGrowth`, and `?ToothGrowth`.

a. Fit a linear regression model to predict tooth length based on the predictors `dose` and the supply delivery method of vitamin C (`supp` where Orange Juice (OJ) = 0 and Ascorbic Acid (VC) = 1) including all interactions.

1. Is delivery method significant?
2. Is there a significant interaction between the `dose` and delivery method?
3. Write two regression equations explicitly, one equation for each delivery method.
4. What percent of the total variation of the tooth length is explained by this regression?

b. Conduct a lack-of-fit test to decide whether the relation between the dose and the tooth length is linear or their might be a better non-linear relationship.

- State the test statistic, the p-value, and your conclusion.

c. Use `plot()` to examine the studentized residuals. Are there any obvious extreme values? Explain your reasoning.

d. Create training and testing subsets and estimate the prediction mean squared error. Use a seed of 1234 and a 60% split.