36-402 – Old Midterm Exam Two

Name ______________________________

Email ______________________________

Instructions:

1. You have 80 minutes to work on the exam. You will need to stop at 11:50, regardless of the time at which you started.
2. If you circle several answers to multiple-choice questions, the question will be marked incorrect.
3. For questions that can have partial credit, please keep written answers brief and clear to save time. Do not feel pressure to fill the provided text lines or to write complicated sentences.
4. Illegible work will not receive any credit.
5. You are allowed a non-graphic calculator, but not your phone. You are also allowed one page (8.5" by 11" sheet of paper) of handwritten notes with writing on two sides.
6. You are provided with a table of named distributions (last page).
7. If you don't have a calculator, answers can involve any of the following notation:

$$\begin{pmatrix} n \\ k \end{pmatrix},\ n^k,\ \exp(k), \log(k),\ n!$$

If you would like to receive this exam back during lecture **please sign below** allowing me to distribute it via envelopes that will be passed around the room. **Please note** that the accumulated scores will not be written anywhere on the returned exam. Total scores will be posted on Blackboard.

RELEASE STATEMENT REGARDING RETURN OF EXAMS
"I hereby authorize my professor in 36-402 to distribute my exam via a large envelope which will be passed around the room during lecture. In so doing, I recognize that the exams are not perfectly secure, and I accept the potential that this may allow access to these documents by others."

______________________________ (Sign)

______________________________ (Date)

Page left blank.

Part I: Multiple Choice (Answer Only)

For each of the following questions, simply give the answer (CIRCLE ONE ONLY). It is not required that any derivation or justification be provided, and there is no partial credit. Each question is worth two points.

1. Given X, Y, with $Y = r(X) + \varepsilon$. Suppose that we are trying to predict Y at some fixed value $X = x$, and suppose further that $\text{Var}(\varepsilon | X = x) = \sigma^2$. Then the expected test error (in terms of squared error loss)

   (a) Must be $\geq \sigma^2$

   (b) Can be $< \sigma^2$

2. Suppose we regress Y on X and Z. Which of the following is the same as the regression $r(x) = \mathbb{E}[Y|X = x]$? Circle one.

   (a) $r_A(x) = \int \mathbb{E}(Y|Z = z, X = x) f(z) dz$

   (b) $r_B(x) = \int \mathbb{E}(Y|Z = z, X = x) f(z|x) dz$

3. The smallest degrees of freedom that a kernel smoother can possibly achieve (as we vary its bandwidth parameter h) is

   (a) 0

   (b) 1

   (c) 2

   (d) 3

4. Given a sample of size n, leave-one-out cross-validation is the same thing as

   (a) 1-fold cross-validation

   (b) 2-fold cross-validation

   (c) $(n-1)$-fold cross-validation

   (d) n-fold cross-validation

   (e) None of the above

5. Degrees of freedom

   (a) Measures the effective number of parameters used by an estimator

   (b) Is random, i.e., it depends on the data

   (c) Is always an integer

   (d) Cannot be negative

   (e) Both (a) and (c)

   (f) Both (a), (c), and (d)

   (g) All of the above

6. Suppose that we observe (x_i, Y_i), $i = 1, \dots n$, where the inputs x_i, $i = 1, \dots n$ are considered fixed, and $Y_i = r(x_i) + \epsilon_i$, where ϵ_i, $i = 1, \dots n$ are i.i.d. with mean 0 and variance σ^2. Let $\widehat{r}$ be an be an estimator of the underlying regression function r, that has 0 expected training error. Then the expected test error is equal to

   (a) $\mathrm{df}(\widehat{r})/n$

   (b) $\sigma^2 \mathrm{df}(\widehat{r})/n$

   (c) $2\sigma^2 \mathrm{df}(\widehat{r})/n$

   (d) None of the above

7. An additive model

   (a) Generally suffers from high variance but low bias, compared to an unrestricted nonparametric estimator of p predictor variables

   (b) Allows us to estimate a nonparametric regression function of p predictor variables, using tools from univariate nonparametric regression

   (c) Computationally requires us to compute only p univariate nonparametric fits—one for each predictor variable

   (d) Both (b) and (c)

8. A generalized linear model

   (a) is usually not parametric but a linear model (that is, a special case of a generalized linear model) is

   (b) is neither parametric or fully nonparametric

   (c) both (a) and (b)

   (d) none of the above

9. Iteratively reweighted least squares is useful because

   (a) It simplifies the computations by assuming that the variance of Y_i is constant in a generalized linear model

   (b) It allows us to calculate the maximum likelihood estimate for the coefficients in a generalized linear model

   (c) It gives us a way of forming hypothesis tests and confidence intervals for coefficients in a generalized linear model

   (d) Both (b) and (c)

   (e) All of the above

10. Bootstrap methods rely on less assumptions than built-in R functions so they always lead to more accurate results but at a higher computational cost

   (a) TRUE

   (b) FALSE

Part II: Multiple Choice with Brief Justification

Each multiple choice question below is worth 3 points. One point is awarded for selecting the correct answer (**circle one only**). Another two points are awarded for correctly justifying your answer. Your justification can be brief.

1. Suppose $Y = \beta_1 X + \beta_2 Z + \varepsilon$, where $\varepsilon \sim N(0, 1)$, and X and Z are normally distributed random variables. Regress Y on X only (with Z omitted). Will this give you a linear regression with the same coefficient β_1? Circle the right answer.

(a) Yes, always.

(b) No, never.

(c) It depends.

Provide your justification:

2. Suppose you use K-fold cross-validation (CV) to estimate the prediction error in regression. Which of the following CV approaches has the smallest bias? Circle one.

(a) 2-fold cross-validation, i.e. K=2

(a) 5-fold cross-validation, i.e. K=5

(c) 10-fold cross-validation, i.e. K=10

(d) The bias in the error estimate is the same regardless of K.

Provide your justification:

Part III: Short Exercises and Short Answer Questions

Please keep written answers brief and clear to save time. Do not feel pressure to fill the provided text lines or to write complicated sentences.

1. Consider the prediction error $R(h) = \mathbb{E}[(Y - \widehat{r}_n(x))^2]$ of a regression kernel smoother $\widehat{r}_n(x)$ with bandwidth h. With big-O notation, we have for $x \in \mathbb{R}$ the general form:

$$R(h) = O(h^4) + O\left(\frac{1}{nh}\right) + \sigma^2, \tag{1}$$

where σ^2 is a positive constant. To minimize $R(h)$, we need to choose the bandwidth according to $h_* = O(n^{-\alpha})$ which will give the optimal nonparametric rate $R(h_*) = O(n^{-\beta})$.

(a) (3 pts) Derive the values of α and β, by balancing the bias and variance terms as in lecture.

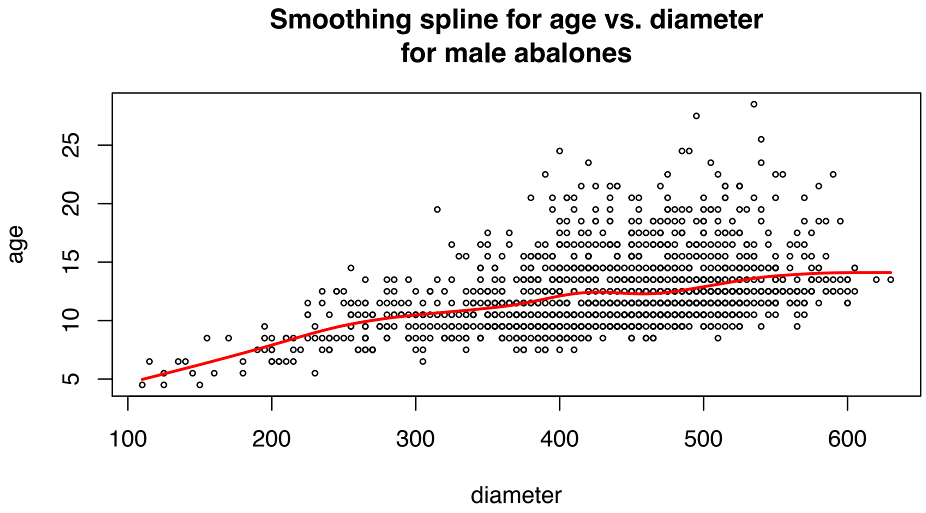
(b) (3 pts) Suppose now that $x \in \mathbb{R}^p$ and that we use a (spherically symmetric) multivariate kernel regression smoother with bandwidth h. The prediction error then has the general form:

$$R(h) = O(h^4) + O\left(\frac{1}{nh^\gamma}\right) + \sigma^2. \tag{2}$$

What is the value of γ? Explain.

(c) (5 pts) For part (a), draw a plot of generalization error $R(h)$ versus h, where you schematically show the decomposition of $R(h)$ into the 3 terms in Eq.1; that is, plot one curve for each of the three term as well as their sum $R(h)$ as a function of h (for fixed n). Mark the values h_* and $R(h_*)$ in your plot.

(d) (5 pts) How do the 4 curves in part (c) change as you *increase* n to a larger value? Draw a figure that illustrates this schematically. (The figure only needs to be clear enough to correctly show for each curve: if there is a change, and if so, whether the curve then shifts left-right and/or up-down.) Mark the new values of h_* and $R(h_*)$ in your plot.



2. As part of the last midterm, you fitted a linear model, as well as a smoothing spline model (see figure above) of age versus diameter for male abalones.

(a) (3 pts) Suppose you would like to compute an (approximate) point-wise confidence band for the true regression function and that you use bootstrap for that. Based on the figure above, which bootstrap approach (parametric, residual, pairs) would you use and why?

(b) (3 pts) Briefly describe how you would create a bootstrap sample for part (a); that is, how you sample X_i, and how you sample Y_i and what step(s) you repeat. You can write "pseudo-code" or bullets rather than full sentences. (You don't need to write down the expression for how you compute the bootstrap confidence interval.)

(c) (5 pts) Suppose you use bootstrap to check the regression specifications of your *linear* model. Describe how you would go about this: What are your null and alternative hypotheses? What is your test statistic? How do you create a bootstrap sample and simulate the appropriate sampling distribution? How would you decide whether the model fits the data?

3. These data [Koch & Edwards (1988)] are from a double-blind clinical trial investigating a new treatment for rheumatoid arthritis. The subjects are randomly assigned to treatment or placebo. The data frame has 84 observations and 5 variables. To fit this model in R we use the `glm` command, which stands for **generalized linear model**.

- ID: patient ID.
- Treatment: factor indicating treatment (Placebo, Treated).
- Sex: factor indicating sex (Female, Male).
- Age: age of patient.
- Improved: ordered factor indicating treatment outcome (None, Some, Marked).

To display and model the data, we try the following:

```
> library(vcd)
> data(Arthritis)
> art <- xtabs(~ Treatment + Improved, data = Arthritis,
+ subset = Sex == "Female")  # to display data in a table
> art
         Improved
Treatment None Some Marked
  Placebo   19    7      6
  Treated    6    5     16

> mosaic(art, gp = shading_max) #displays this information in a figure

#Create a binary response from the 3-level response in original data

> response = 1- (Arthritis$Improved == "None")

> mylogit<- glm(response ~ Treatment + Sex + Age,
+ family=binomial, data = Arthritis)
> summary(mylogit)

Call:
glm(formula = response ~ Treatment + Sex + Age, family = binomial,
    data = Arthritis)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.10833  -0.91158   0.05362   0.91681   1.84659

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.01546    1.16777  -2.582  0.00982 **
TreatmentTreated  1.75980    0.53650   3.280  0.00104 **
```

```
SexMale          -1.48783    0.59477  -2.502  0.01237 *
Age               0.04875    0.02066   2.359  0.01832 *
---

    Null deviance: 116.449  on 83  degrees of freedom
Residual deviance:  92.063  on 80  degrees of freedom
AIC: 100.06

> confint(mylogit) #confidence intervals for coefficients.
                        2.5 %      97.5 %
(Intercept)      -5.477304188 -0.84351926
TreatmentTreated  0.750803551  2.87506538
SexMale          -2.729719456 -0.37239953
Age               0.009951561  0.09194283
```

Answer the following questions using the R output:

(a) (2 pts) How does treatment affect your odds of improvement (versus none)?

(b) (2 pts) How does gender affect the odds of improvement (versus none)?

(c) (2 pts) Can you draw causal inferences about the treatment from this study? Why or why not?

(d) (3 pts) Provide an approximate 95% confidence interval (CI) for how age affects the odds of improvement. Make sure that it is clear what population quantity the CI is for; that is, the final answer should have the form "a 95% CI for [fill in] is [fill in]".

(e) (5 pts) Briefly explain how you could use the R output above to test how well your model fit the data (globally): What is your null versus alternative hypotheses? What is your test statistic? When would you reject the null at, say, level $\alpha = 0.05$?

4. (5 pts) In one of your homework assignments, you predicted the death rate in Chicago by fitting an *additive model for the log of the death counts* on various predictors $X_1, \ldots, X_p$ for pollution and temperature. How is this any different from a *Poisson-response GAM*?
(Note: For full credit, you should for each case, explain what the regression model is, including what the response variable Y is, and what the probability model of Y given $X = (x_1, \ldots, x_p)$ is.)