NAME (PRINT):

Last/Surname First /Given Name

STUDENT #: **SIGNATURE:**

UNIVERSITY OF TORONTO MISSISSAUGA
DECEMBER 2018 FINAL EXAMINATION
STA314H5F
Introduction to Statistical Learning
Stanislav Volgushev
Duration - 2 hours
Aids: None

The University of Toronto Mississauga and you, as a student, share a commitment to academic integrity. You are reminded that you may be charged with an academic offence for possessing any unauthorized aids during the writing of an exam. Clear, sealable, plastic bags have been provided for all electronic devices with storage, including but not limited to: cell phones, SMART devices, tablets, laptops, calculators, and MP3 players. Please turn off all devices, seal them in the bag provided, and place the bag under your desk for the duration of the examination. You will not be able to touch the bag or its contents until the exam is over.

If, during an exam, any of these items are found on your person or in the area of your desk other than in the clear, sealable, plastic bag, you may be charged with an academic offence. A typical penalty for an academic offence may cause you to fail the course.

*Please note, once this exam has begun, you **CANNOT** re-write it.*

Marks possible: 45 **Marks achieved**:

Problem	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Possible	2	1	1.5	1	1	1	1	1.5	3	4	2.5	1.5	1	3	3	4	1	1	1	1
Achieved																				

Problem	21	22	23	24	25
Possible	1	1	1	4	2
Achieved					

1. (0.5 marks each) Assume that `D` is a data frame that contains 4 columns with names `Y`, `X`, `V`, `W`. For each of the following specifications, describe the regression function that corresponds to the `lm` call in **R**. (you may use terms such as step function, spline, piecewise polynomial, local polynomial regression, smoothing spline etc.)

   (a) `lm( Y ~ . -V - W, data = D)`
   $f(x, v, w) = b_1 + b_2 x$, a linear function in x

   (b) `lm( Y ~ V + X + V:W, data = D)`
   $f(x, v, w) = b_1 + b_2 v + b_3 x + b_4 vw$

   (c) `lm( Y ~ X + bs(W,knots=c(1,3),degree = 2), data = D)`
   An additive model with linear effect of X and effect of W described by a polynomial spline of degree 2 with knots in $1, 3$.

   (d) `lm( Y ~ ns(X,knots = c(2,5)) + I(V^3), data = D)`
   An additive model with terms V^3 and effect of W described by a natural cubic spline with knots in $2, 5$.

2. (0.5 marks each) Assume that `D` is a data frame that contains 4 columns with names `Y`, `X`, `V`, `Z`. For each of the following regression functions, decide if they can be formulated as a linear regression in **R** (here, $b_1, ..., b_3$ are unknown). If yes, write the **R** call you would use.

   (a) $f(x, v, z) = b_1 + b_2 z + b_3 vz$

   `lm( Y ~ Z + V:Z, data = D)`

   (b) $f(x, v, z) = b_1 + b_2\sqrt{z + b_3 v}$
   Not possible

3. *Running a linear regression in* ***R*** *and applying the summary function you get the following output*

```
Call:
lm(formula = y ~ x1 + x2 + x1:x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29881 -0.07232 -0.01262  0.07602  0.28263

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.100000   0.011326  -1.048    0.297
x1           0.900000   0.011910  83.384   <2e-16 ***
x2          -0.050000   0.011095  -0.762    0.448
x1:x2        0.100000   0.011635  0.359     0.672
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.111 on 96 degrees of freedom
Multiple R-squared:  0.7929,    Adjusted R-squared:  0.7927
F-statistic:  4487 on 3 and 96 DF,  p-value: < 2.2e-16
```

(a) (1 mark) Given the output above, what is your prediction for a new observation with predictor values `x1 = 1, x2 = 2`? You don't need to simplify your answer, it is enough if you write down the correct formula.

Solution $-1.1 + 0.9 * 1 - 0.05 * 2 + 0.1 * 1 * 2$

(b) (0.5 marks) What is the value for R^2 in the model above?
0.7929

4. (1 mark) Which output will running the following code in **R** give?

```
x = array(0,3)
for(k in 1:length(x)){
 x[k] = k-3
}
which.min(x)
```

5. (1 mark) Which output will running the following code in **R** give (you don't need to simplify the answer, writing down the correct formula is enough)?

```
x = array(0,3)
for(k in 1:3){
 x[k] = k^2
}
x[-1]
```

6. (1 mark) Assume that you observe data (x_i, y_i) with values $(1,2), (2,3), (4,5), (0,2), (3,7)$. Compute the 3-nn estimator for $x = 10$.

   $(3 + 5 + 7)/3$

7. (1 mark) You run 12-fold cross-validation on a data set and obtain the following values for the cross-validated error and standard error for different numbers d of predictors in linear models selected by best subset selection.

d	0	1	2	3	4	5	6	7	8	9	10
cv(d)	13	12	7	6	2	3	4	1	6	15	30
$12^{-1/2}\widehat{se}(d)$	8	5	4	3	3	5	6	7	5	6	5

   Which d would you select based on the one standard error rule?

   $d = 2$

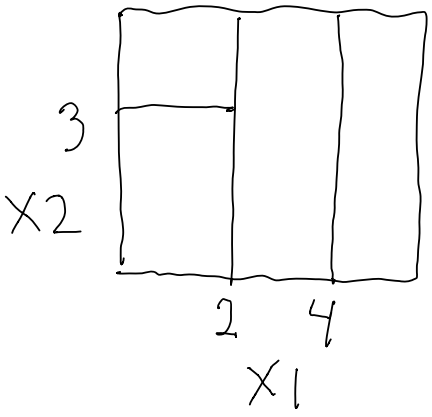
8. (1.5 marks) You run 3-fold cross-validation on a data set and obtain the following values for the cross-validated error and standard error for different values of m (number of principal components) in PCR.

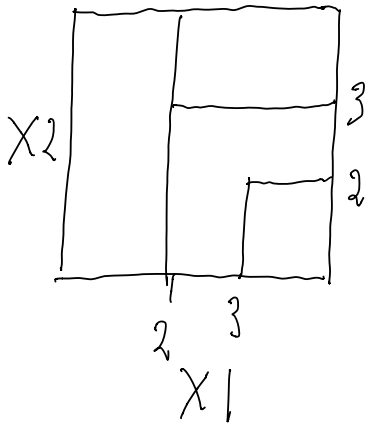
m	1	2	3	4	5	6	7	8	9	10
cv(m)	12	7	6	X	3	4	1	6	15	30
$3^{-1/2}\widehat{se}(m)$	15	14	13	13	8	6	7	5	16	15

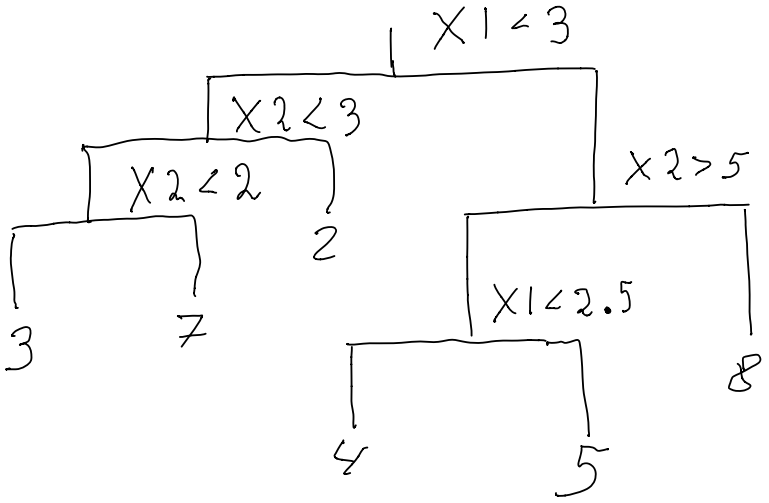
Is it possible to find a value for X such that $m = 1$ is selected by the one standard error rule in cross-validation? Justify your answer!

Answer: not possible

9. (1.5 marks each) For each of the two partitions of the predictor space given below, is it possible to represent them as a regression tree? If yes, draw the corresponding tree.







10. *For the regression tree shown above*

    (a) (1 mark) What is the predicted value for a new point with predictor X = (2, 1)?

    (b) (1.5 marks) Draw the corresponding partition of the predictor space next to the tree.

    (c) (1.5 marks) Assume that X = (1,$*$). Is it possible to find a value for $*$ so that the predicted value for a new point is 5? Justify your answer!

11. Consider a linear model with 3 predictors, A, B, C. The following table gives the RSS for each combination of predictors in a linear model.

none	A	B	C	A,B	A,C	B,C	A,B,C
7	3	4	5	2	2.5	1.5	1

(a) (0.5 marks) Which model with 2 predictors is selected by forward stepwise selection?

A,B

(b) (1 mark) Assume that additionally $\hat{\sigma}^2 = 0.6$ and $\log n = 2$. Which model will be selected by best subset selection with C_p?

B,C

(c) (1 mark) Assume that additionally $\hat{\sigma}^2 = 0.1$ and $\log n = 5$. Which model will be selected by backward stepwise selection with BIC?

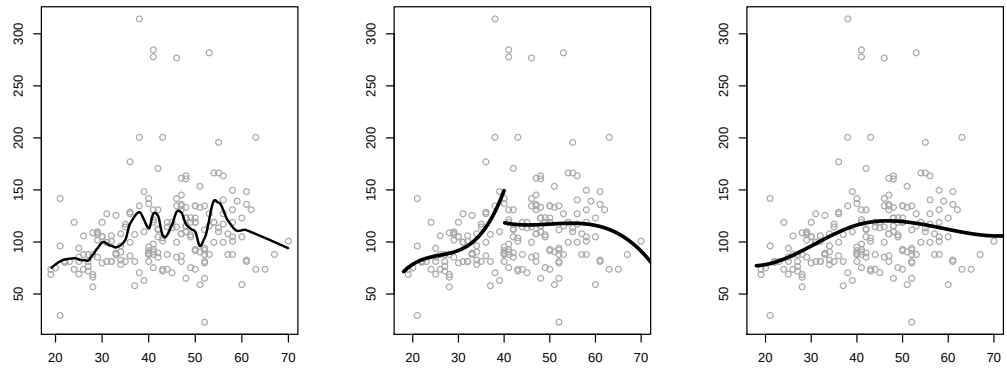
A,B,C or B,C (either answer would have been correct)

12. (1.5 marks) Assume you have 7 predictors. Will best subset selection with AIC and backward stepwise selection with AIC always result in the same model with 6 predictors? Justify your answer.

Answer: yes.

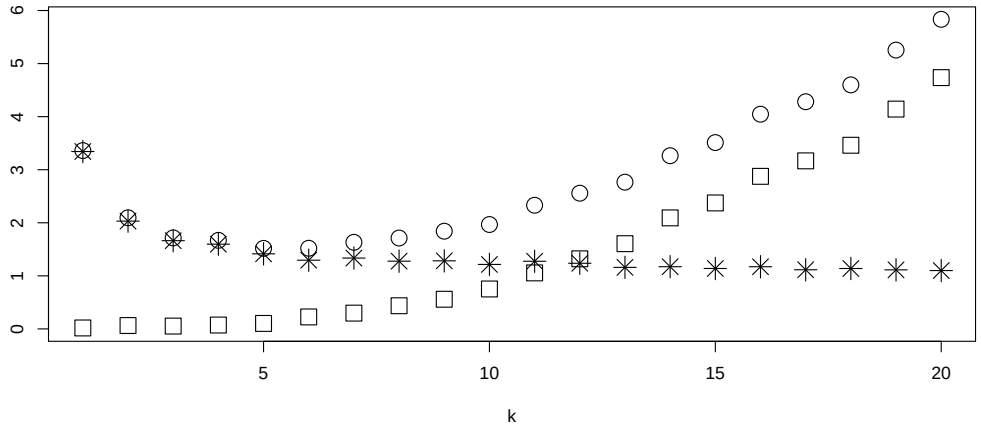
13. (1 mark) In the notation given in lectures, assume that $\tilde{X}^T\tilde{X}$ has eigenvectors v_1, v_2, v_3 corresponding to eigenvalues $2, 1.5, 3$. What are the first and second principal directions (principal components) in PCR?

First: v_3. Second: v_1.



14. (3 marks) The plots at the top of this page show the graphs of 3 different regression functions: a spline of degree 3 with 1 knot, a spline of degree 3 with 20 knots, and a piecewise polynomial of degree 3 with one knot. All functions were learned from the displayed data set by least squares. Which plot is which? Justify your answer.

    Answer: left plot: spline of degree 3 with 30 knots, middle plot: piecewise polynomial, right plot: spline with one knot.



15. (3 marks) The plot on top of this page shows squared bias, variance, and MSE of k-nn on a test set as a function of k. Which one is which? Justify your answer!

    Answer: circles are MSE, squares are squared bias, stars are variance.

16. Assume you want to model the influence of a predictor `x` on a response `y` by the regression function f which is a cubic spline (i.e. polynomial spline of degree 3) with a knot in the point 3.

(a) (1 mark) How many degrees of freedom (**counting the way we counted in lectures**) does this model have?
5

(b) (1 mark) Assume that `D` is data frame with columns `y` (response) and `x` (predictor). Which `R` input would you use to fit this model?

```
lm( Y ~ X + bs(W,knots=c(1,3),degree = 2), data = D)
```

(c) (2 marks) Write down functions $g_1, g_2, ..., g_d$ such that f is a a cubic spline with knots in the points 0 and 1 if and only if

$$f(x) = b_1 + b_2 g_1(x) + ... + b_{d+1} g_d(x)$$

for some $b_1, ..., b_{d+1} \in R$.
$g_1(x) = x, g_2(x) = x^2, g_3(x) = x^3, g_4(x) = (x-0)_+^3, g_5(x) = (x-1)_+^3$

17. (**1 mark if and only if ticks are only next to correct answers. Any number of answers (0-3) can be correct**) Which of the following methods can still give reasonable predictions when the number of predictors $p = 567$ is larger than the number of observations $n = 100$?

- ☐ PCR with number of principal components $m > n$.
- ☐ An additive model with polynomials of degree 3 for each predictor.
- ☐ Ridge regression with $\lambda = 0$.

18. (**1 mark if and only if ticks are only next to correct answers. Any number of answers (0-3) can be correct**) Compared to growing a large single tree, using bagging

- ☐ Can be better because it helps to reduce bias.
- X Can be better because it helps to reduce the variance.
- ☐ Will only be better if the number of trees is small.

19. (**1 mark if and only if ticks are only next to correct answers. Any number of answers (0-3) can be correct**) Using 5-fold cross validation with a data set of size $n = 30$ to select a tuning parameter

- X Will usually be computationally less expensive than leave-one-out cross-validation on the same data set.
- X Might give different answers since some randomness is involved.
- ☐ Does not make sense since $n > 5$.

20. (**1 mark if and only if ticks are only next to correct answers. Any number of answers (0-3) can be correct**) Which of the following changes correspond to *increasing* the flexibility of a model (usually this corresponds to decreasing the training error)

- ☐ Removing predictors from a linear model.
- ☐ Increasing λ in lasso.
- X Decreasing the span in local regression.
- ☐ Pruning back large regression trees.

21. (**1 mark if and only if ticks are only next to correct answers. Any number of answers (0-3) can be correct**) Interaction effects between the predictors x_1, x_2 in linear regression

- X Can be incorporated by including a term of the form bx_1x_2.
- ☐ Can be incorporated by including a term of the form $b(x_1 + x_2)$.
- ☐ Can not be incorporated since the model would become nonlinear.

22. (**1 mark if and only if ticks are only next to correct answers. Any number of answers (0-3) can be correct**) A high value of R^2 (a value close to one) in a linear regression model

- ☐ Means that the regression model is incorrect.
- ☐ Means that there is no relationship between the predictors and the response.
- ☐ Means that using the predictors to construct a linear model leads to better predictions than ignoring the predictors. **This is false because R^2 can be large even if there are many irrelevant predictors**

23. (**1 mark if and only if ticks are only next to correct answers. Any number of answers (0-3) can be correct**) Increasing the parameter λ in smoothing spline regression

- X Can increase the training error.
- X Will never decrease the training error.
- ☐ Will never affect the training error.

24. (4 marks) Consider a linear regression with one-dimensional predictor x and data $(x_i, y_i)_{i=1,\dots,n}$. Assume that the predictor is standardized such that $\bar{x} := \frac{1}{n}\sum_{i=1}^n x_i = 0$.

(a) (1 mark) Write down the minimization problem that correspond to ridge regression with parameter λ (recall: the intercept is not penalized).
See lectures.

(b) (3 marks) Compute the explicit form of the corresponding estimators $\hat{b}_1^R, \hat{b}_2^R$ from scratch (i.e. without using any results from lectures). The following notation might be useful:

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^n x_i y_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^n y_i, \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2$$

See derivations in lectures.

25. (2 marks) Write down the definition of RSS and TSS in a linear model with a one-dimensional predictor x (using formulas). Prove that $TSS \geq RSS(\hat{b})$ where $\hat{b}$ is the least squares estimator (the estimator we discussed in class) in the linear model **without using properties of R^2 that were derived in lectures.**

See derivations in lectures.

End of Exam