MS&E 125 Final (sample)

Name: _____

SUNet ID: _____@stanford.edu

Instructions

- Unless otherwise noted, each problem is self contained.
- Clearly mark your answer on the ***answer sheet***, found on the last page.
- *Only the answer sheet will be graded*, with no partial credit.
- Questions with more than one answer marked on the answer sheet will be considered incorrect.
- There are a total of 40 questions, 10 true/false questions and 30 multiple choice questions.
- Each problem is identically worth 2.5 points, for a maximum total of 100, so don't spend too much time struggling on any single question.
- You will get zero points for any incorrectly answered question (no negative points).
- You may use all resources available (books, notes, homework solutions, and general Internet) for this exam. However, we recommend against using materials outside of this course, such as searching the Internet, since it will more likely result in over-complication, confusion and a waste of time.
- Good luck!

True/false questions

Problem 1.
For independent samples $X_1, X_2, \ldots, X_n$ from an exponential distribution with parameter θ, the sampling distribution of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately normal for large n.

Problem 2.
For a fixed set of covariates, a fitted L^2 regularized (ridge) regression model will typically have more non-zero coefficients than an L^1 regularized (lasso) regression model.

Problem 3.
Let $X_1, ..., X_n$ represent samples taken from a Bernoulli(p) distribution. Consider the sample mean estimator $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. As the number of samples n increases, the bias of $\hat{p}$ gets smaller.

Problem 4.
As the number of samples n increases, the standard error of $\hat{p}$ in problem 3 gets smaller.

Problem 5.
Jerry and Camelia are both given the same set of data, which they use to compute a bootstrap standard error using 10,000 bootstrap samples for some estimator $\hat{\theta}$. Because the data are identical, Jerry and Camelia will compute the exact same value for their standard error.

Problem 6.
Consider a linear regression model. The width of a prediction interval for a **specific** response will typically be wider than the width of the confidence interval for the **mean** response.

Problem 7.
You are given a dataset of two variables: `hon` and `female`, where `hon` is a binary variable indicating whether or not a student is in an honors class, and `female` is a binary variable indicating whether or not the student is female.

You train the following logistic regression model:

```
model <- glm(hon ~ 1 + female, family = binomial, data = honor)
```

These are the coefficients that define your trained model.

```
##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -1.4708517  0.2689554 -5.468756 4.532047e-08
## female       0.5927822  0.3414293  1.736178 8.253231e-02
```

In this dataset, men are more likely than women to be in an honors class.

Problem 8.
For an (unregularized) logistic regression fit on some training set, the AUC on a test set will typically be lower than the AUC on the training set.

Multiple choice questions

As a data scientist at Webflix, you are given views data. Each row in the dataset indicates views of a movie for a given user during the first quarter of 2019. There are 70 movies and 10 users, for a total of 700 rows. You load the data in `R` as `views_df` and see that it has the following columns:

- `user`: a character column of unique user IDs for a total of 10 users
- `movie`: a character column of 70 unique movie IDs
- `viewed`: a logical column of `TRUE/FALSE` indicating whether each user watched the corresponding movie during the first quarter of 2019.

Answer problems 9 and 10.

Problem 9.
We would like to know how many users watched each movie. What are the correct values for each of the blanks in the following snippet

```
views_df %>%
  group_by([  A  ]) %>%
  summarize(total_views = [  B  ](viewed))
```

(a) `[ A ]`: `user`, `[ B ]`: `sum`

(b) `[ A ]`: `user`, `[ B ]`: `max`

(c) `[ A ]`: `movie`, `[ B ]`: `sum`

(d) `[ A ]`: `movie`, `[ B ]`: `max`

Problem 10.
How many rows will the following code snippet result in?

```
views_df %>%
  group_by(movie) %>%
  summarize(p_view = mean(viewed)) %>%
  filter(p_view > .6)
```

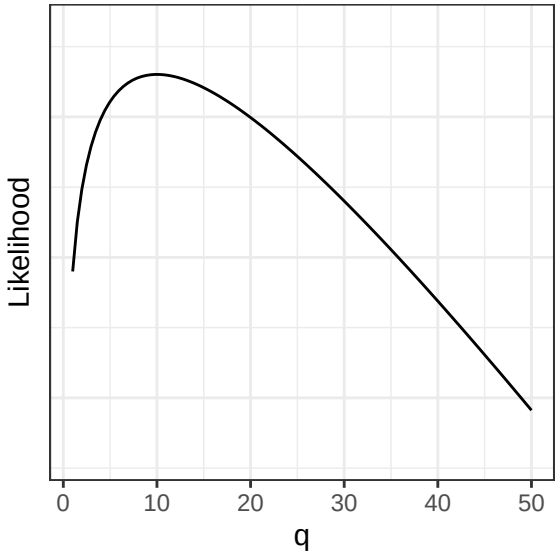
(a) 10

(b) 70

(c) 700

(d) Cannot determine without running the code

Problem 11.
The figure below shows the likelihood of some observed data X as a function of *all* possible values for the parameter q.



Which of the following statements is true?

(a) The MLE for q is 10

(b) The MLE for q is $\min(X)$

(c) The MLE for q is $\max(X)$

(d) The MLE for q cannot be determined from the plot

Problem 12.
You are given the model, `house_price ~ 1 + year_built + sq_feet + school_district_quality`, where `year_built` and `sq_feet` are continuous covariates, and `school_district_quality` is a categorical variable expressing the quality of the school in the area, with four possibe values: `A`, `B`, `C`, and `D`. How many coefficients does the model have (including the intercept)?

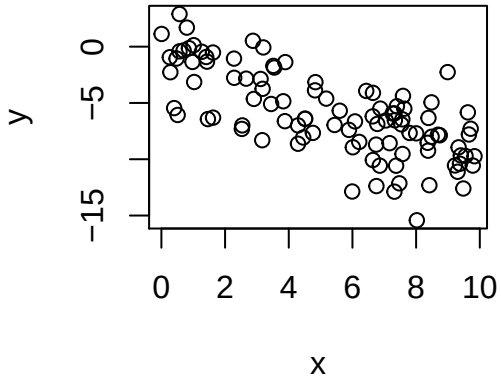
(a) 7

(b) 6

(c) 4

(d) 3

Problem 13.
Included below is a plot of two variables x and y.



Let ρ denote the correlation coefficient between x and y. Which of the following best describes the value of ρ?

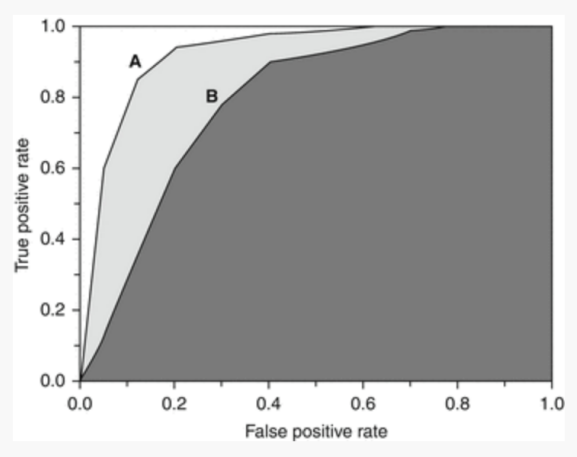
(a) $\rho = -1$

(b) $-1 < \rho < 0$

(c) $\rho = 0$

(d) $0 < \rho < 1$

Problem 14.
Consider the following plot showing the ROC curves for two different logistic regression models A and B.



Which of the following is a valid conclusion to draw about the performance (as measured by AUC) of these two models?

(a) Model A has better performance than model B

(b) Model B has better performance than model A

(c) Models A and B have fairly comparable performance

(d) There isn't enough information in the plot to draw meaningful conclusions about the performance of A and B

Problem 15.
Which of the following statements about model selection is NOT true?

(a) Regularization can help reduce overfitting

(b) The train-validate-test process can help detect model overfitting

(c) Cross validation is particularly helpful when you have a lot of training data

(d) Maximizing performance on a training set can lead to overfitting

Problem 16.
Researchers are planning to conduct a study of the effect of an occupational exposure on health outcomes. The researchers plan to study exposed workers from one factory and compare them with unexposed retirees who have never worked in a factory. A reviewer of the research proposal is worried about selection bias. Which of the following best represents the reviewer's concern?

(a) Retirees should not be compared to factory workers because factory workers are under more stress than retirees

(b) Retirees should not be compared to factory workers because factory workers' incomes differ from those of retirees

(c) Retirees should not be compared to factory workers because factory workers are likely to need to maintain a certain level of health in order to work in a factory while retirees would not necessarily be as healthy

(d) Retirees should not be compared to factory workers because factory workers likely live in a different city than the retirees

Answer sheet

Name: _____

SUNet ID: _____@stanford.edu

True/false questions

Fill-in the circle of the correct answer. (T = true, F = false)

1	Ⓣ	Ⓕ	**6**	Ⓣ	Ⓕ
2	Ⓣ	Ⓕ	**7**	Ⓣ	Ⓕ
3	Ⓣ	Ⓕ	**8**	Ⓣ	Ⓕ
4	Ⓣ	Ⓕ	**9**	Ⓣ	Ⓕ
5	Ⓣ	Ⓕ	**10**	Ⓣ	Ⓕ

Multiple choice questions

11	ⓐ	ⓑ	ⓒ	ⓓ	**21**	ⓐ	ⓑ	ⓒ	ⓓ	**31**	ⓐ	ⓑ	ⓒ	ⓓ
12	ⓐ	ⓑ	ⓒ	ⓓ	**22**	ⓐ	ⓑ	ⓒ	ⓓ	**32**	ⓐ	ⓑ	ⓒ	ⓓ
13	ⓐ	ⓑ	ⓒ	ⓓ	**23**	ⓐ	ⓑ	ⓒ	ⓓ	**33**	ⓐ	ⓑ	ⓒ	ⓓ
14	ⓐ	ⓑ	ⓒ	ⓓ	**24**	ⓐ	ⓑ	ⓒ	ⓓ	**34**	ⓐ	ⓑ	ⓒ	ⓓ
15	ⓐ	ⓑ	ⓒ	ⓓ	**25**	ⓐ	ⓑ	ⓒ	ⓓ	**35**	ⓐ	ⓑ	ⓒ	ⓓ
16	ⓐ	ⓑ	ⓒ	ⓓ	**26**	ⓐ	ⓑ	ⓒ	ⓓ	**36**	ⓐ	ⓑ	ⓒ	ⓓ
17	ⓐ	ⓑ	ⓒ	ⓓ	**27**	ⓐ	ⓑ	ⓒ	ⓓ	**37**	ⓐ	ⓑ	ⓒ	ⓓ
18	ⓐ	ⓑ	ⓒ	ⓓ	**28**	ⓐ	ⓑ	ⓒ	ⓓ	**38**	ⓐ	ⓑ	ⓒ	ⓓ
19	ⓐ	ⓑ	ⓒ	ⓓ	**29**	ⓐ	ⓑ	ⓒ	ⓓ	**39**	ⓐ	ⓑ	ⓒ	ⓓ
20	ⓐ	ⓑ	ⓒ	ⓓ	**30**	ⓐ	ⓑ	ⓒ	ⓓ	**40**	ⓐ	ⓑ	ⓒ	ⓓ

Solutions

Problem 1: True

Problem 2: True

Problem 3: False

Problem 4: True

Problem 5: False

Problem 6: True

Problem 7: False

Problem 8: True

Problem 9: (c)

Problem 10: (d)

Problem 11: (a)

Problem 12: (b)

Problem 13: (a)

Problem 14: (a)

Problem 15: (c)

Problem 16: (c)