UNIVERSITY OF ILLINOIS, URBANA-CHAMPAIGN
Department of Computer Science **CS 498: Probability and Statistics for CS Undergrads**

Professor: David Forsyth

Final Examination

NAME:————

Fill the answers and your name in on the exam, and return it. You have 180 mins.
There are a total of 71 marks available, and full marks is 71 marks.
Good luck, and be careful.

Important note: by submitting your exam for grading, you are certifying that you referred only to your single page of notes in preparing your answer, and not to any other source of reference, including the papers of other students.

Question	Marks	Out of
Hypotheses and populations		7
Hypotheses and populations		10
Models and populations		11
Classification		9
Models and populations		10
Principal Components		10
Linear Regression		2
Linear Regression		6
Linear Regression		6

Hypotheses and Populations

You wish to estimate the mean of a population. Although you do not know it, the population has mean 10, and standard deviation 3.

- Assume that you know that the population has standard deviation 3. You draw 1 sample randomly from the population.

    1. How can you estimate the population mean from this sample? (2)

    2. How accurate is your estimate? (i.e. what is its standard deviation) (2)

    3. How can you make your estimate have standard deviation 0.1? (2)

- Assume that you do not know the population standard deviation. How can you estimate it? (1)

Hypotheses and Populations

I have two populations. I draw a random sample from population A, which has 10 items in it. The mean of this sample is 0. The standard deviation is 1. I now draw a random sample from population B, which has 10 items in it. The mean of this sample is 1, and the standard deviation is 0.5.

- The mean of the sample from population A is a random variable.
    1. What probability distribution does it have? (1)
    2. What is the mean of that probability distribution? (1)
    3. Give an estimate of the standard deviation of that probability distribution? (1)
- Now assume the hypothesis that populations A and B are the same. In this case, the difference between the sample means is a random variable.
    1. What probability distribution does it have? (1)
    2. What is the mean of that probability distribution? (1)
    3. Give an estimate of the standard deviation of that probability distribution. (1)
- Does the evidence support the hypothesis that populations A and B are the same? Why? (4)

Models and Populations

You have a die, but you do not know if it is fair or not. Write p_6 for the probability that the die comes up six. You perform the following experiment to determine p: you take a six-sided die, and roll it. If the number that comes up is six, you stop; otherwise you repeat.

- What is the distribution of the number of times you roll the die? (1)

- What is the expected number of rolls? (3)

- What is the name of the procedure to estimate p from this information? (1)

- Assume the die comes up six on the 10'th roll. What is your best estimate of p_6 from this experiment? (3)

- Assume the die comes up six on the 1st roll. What is your best estimate of p_6 from this experiment? (3)

Classification

I wish to classify a high dimensional dataset.

- Why can I not build a histogram based classifier? (3)

- How does Naive Bayes help? (3)

- What assumption does Naive Bayes require? (3)

Models and Populations

I assume that newspaper reports about the global warming appear with a Poisson distribution.

- How could I estimate the intensity of this distribution, λ? (2)

- Assume you are given a value of λ. How could you test the hypothesis that this value correctly describes that poisson distribution? (3)

- What does the assumption that these reports appear with a Poisson distribution mean (in terms of independence)? (2)

- Assume that there are two newspapers, one strongly right wing and one strongly left wing. In this case, would the Poisson assumption be reasonable? Why do you think so? (3)

Principal Components

I have a set of data points. Each is an N-dimensional vector. Write $\mathbf{x}_i$ for the i'th such vector.

- Write an expression for the mean of this dataset (1)

- Write an expression for the covariance matrix of this dataset (1)

- Show that the covariance matrix is positive semi-definite (i.e. if the matrix is Σ, then for an arbitrary vector $\mathbf{u}$, $\mathbf{u}^T\Sigma\mathbf{u}$ is always non-negative). (3)

- How would I compute the principal components of this dataset? (1)

- Give two uses for these principal components. (4)

Linear Regression

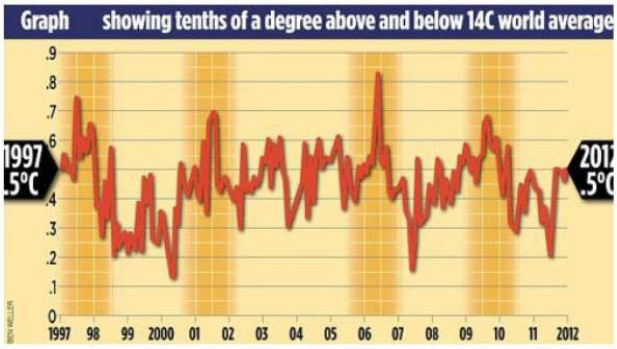


Figure 1:

Figure 1 appeared in newspaper articles during a recent fight over global warming. The figure shows a set of data points of temperature as a function of time for the period 1997-2012.

Assume you use linear regression to fit a model of temperature as a function of time, where T is temperature, t is time. The model has the form $T = \beta_t t + \beta_0$.

1. On the figure, draw the line this model would fit (by eye). (1)

2. Give an approximate value for β_t for this model (by eye). (1)

Linear Regression

We represent a set of N data points using the notation we used in class. We wish to predict temperature from time. We write $\mathbf{Y}$ for a column vector containing all temperature measurements. We write $\mathbf{x}_i = (t_i, 1)^T$ for a column vector containing the time and a 1 for the i'th measurement. We write $\mathcal{X}$ for a matrix containing the data points where

$$\mathcal{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \dots \\ \mathbf{x}_N^T \end{pmatrix}$$

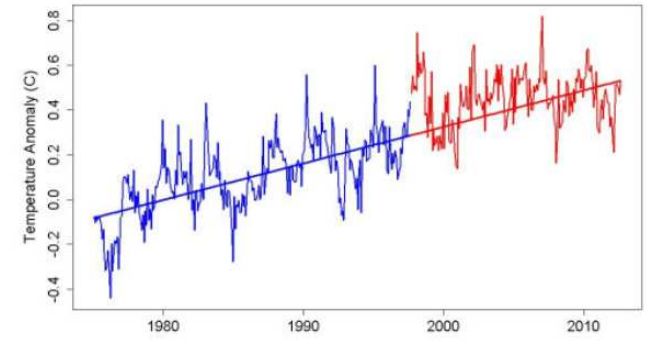
. We model the data as $\mathcal{X}\beta$, and write $\mathbf{e} = \mathbf{Y} - \mathcal{X}\beta$ for the residual error.

1. How do we solve for β? (2)

2. Show that $\mathbf{e}^T\mathcal{X} = 0$, using either an algebraic method or an argument. (3)

3. For this data, $\mathbf{e}^T\mathbf{e}$ is moderately large. What conclusion can we draw? (1)

Linear Regression



Global surface temperatures from the Met Office data. Top: Fiction. Bottom: Fact.
Image credit: David Rose/Daily Mail (top), Tamino (bottom).

Figure 2:

Figure 2 appeared in newspaper articles during a recent fight over global warming. The figure shows a set of data points of temperature as a function of time for the period 1975-2012. Assume you use linear regression to fit a model of temperature as a function of time, where T is temperature, t is time. The model has the form $T = \beta_t t + \beta_0$. The line has been drawn on the figure by eye.

- Give an estimate of the value for β_t for this figure (by eye). (2)

- By eye, choose an interval of 5 or more consecutive years of data where the β_t would be negative. Mark this interval on the graph, and draw the line predicted by regression in this interval (by eye). (2)

- Draw a conclusion about the effects of selecting intervals of data on the models that are fitted. (2)