Chris Piech
CS109

Practice Final
June 5, 2017

Practice Final Examination

This is a closed calculator/computer exam. You are, however, allowed to use notes in the exam. The last page of the exam is a Standard Normal Table, in case you need it. You have 3 hours (180 minutes) to take the exam. The exam is 180 points, meant to roughly correspond to one point per minute of the exam. You may want to use the point allocation for each problem as an indicator for pacing yourself on the exam.

In the event of an incorrect answer, any explanation you provide of how you obtained your answer can potentially allow us to give you partial credit for a problem. For example, describe the distributions and parameter values you used, where appropriate. It is fine for your answers to include summations, products, factorials, exponentials, and combinations, unless the question specifically asks for a numeric quantity or closed form. Where numeric answers are required, the use of fractions is fine.

I acknowledge and accept the letter and spirit of the honor code. I pledge to write more neatly than I have in my entire life:

Signature: ____

Name (print): ____

1. (15 points) Say we have a standard 52 card deck of cards, where the cards have the usual ordering of ranks: 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King, and Ace. Assume each card in the deck is equally likely to be drawn. Two cards are then drawn sequentially (without replacement) from the deck.

   What is the probability that the second card drawn has a rank **greater than** the rank of the first card drawn? Show the derivation you used to compute the probability and also provide your final answer in closed form (if a closed form exists).

2. (20 points) You decide to buy one share of stock in "Happy Computing Inc." (abbreviated: HCI) for \$10. Each day, the price of HCI either increases \$1 (with probability p) or decreases \$1 (with probability $1 - p$). You decide that you will only sell your stock at some point if either it gains \$2 (i.e., reaches a price of \$12) or loses \$2 (i.e., reaches a price of \$8).

   a. (7 points) What is the probability that you will sell your share of HCI (for either a gain or a loss of \$2) *exactly* 4 days after you buy it?

   b. (13 points) Assuming you just bought a share of HCI, so you don't know how long it will be until you sell it, what is the probability (in terms of p) that when you eventually sell it, you will sell it for a gain (i.e., sell it for \$12)? Give a closed form answer.

3. (20 points) Say you are monitoring energy usage for two different types of computers (type W and type X) in an office. There are 10 machines of type W and 10 machines of type X in the office, and the energy usage of all 20 machines are independent.

   Let W_i = the number of kilowatt-hours used by machine i of type W in a day.
   Let X_j = the number of kilowatt-hours used by machine j of type X in a day.

   Say that $W_i \sim \text{Poi}(4)$ for $1 \leq i \leq 10$ and $X_j \sim N(5, 3)$ for $1 \leq j \leq 10$.

   Now say that you have a separate machine that (probabilistically) monitors the energy usage of the 10 type W and 10 type X machines (call these 20 machines the machine pool). Each day, a new "watch list" is generated, where each machine in the machine pool independently has a 0.2 probability of being placed on the "watch list". Let Y = the total energy usage (in kilowatt-hours) of all the machines on the "watch list".

   a. (4 points) What is E[Y]?

   b. (8 points) Say that three type W (and no type X) machines are on the "watch list" on a particular day. What is $P(Y \geq 20)$ that day?

   c. (8 points) Now say that three type X (and no type W) machines are on the "watch list" on a particular day. What is $P(Y \geq 20)$ that day? Give a numeric answer, if possible.

4. (20 points) We have a random number generator that generates random integers independently uniformly distributed from 1 to 5, inclusive. This random number generator is used to generate a sequence on *n* independent random integers.

   Let X = number of 1's generated in the sequence of *n* integers.

   Let Y = number of 5's generated in the sequence of *n* integers.

   (Hint: it might be useful to define indicator variables denoting whether the *i*-th integer generated is a 1, and likewise for whether the *i*-th integer generated is a 5.)

   a. (13 points) What is Cov(X, Y)?

   b. (7 points) What is the correlation ρ(X, Y)?

5. (25 points) Say that our favorite manufacturer of computers produces machines whose lifetimes (independently) each have a mean of 100 weeks with a variance of 25 weeks^2. We buy and use a computer built by this manufacturer until it dies. When the machine dies, we immediately buy a new machine from the same manufacturer, and repeat this process every time the machine we are using dies. Let X = the total number of weeks that we use a machine built by our favorite manufacturer.

   What is the minimal number of machines we would have to buy so that the total number of weeks we use a machine built by our favorite manufacturer is greater than 2000 with 95% probability? Justify your answer mathematically.

   (Hint: derive an expression for the number of machines that will be necessary to purchase in order to have $P(X > 2000) \geq 0.95$, and use that to determine the answer).

6. (25 points) Consider the following recursive functions:

```
int Near() {
   int b = randomInteger(1, 4); // equally likely to be 1, 2, 3 or 4
   if (b == 1) return 2;
   else if (b == 2) return 4;
   else if (b == 3) return (6 + Near());
   else return (8 + Near());
}

int Far() {
   int a = randomInteger(1, 3); // equally likely to be 1, 2 or 3
   if (a == 1) return 2;
   else if (a == 2) return (2 + Near());
   else return (4 + Far());
}
```

Let Y = the value returned by `Far()`.

a. (10 points) What is E[Y]? Give a numeric answer.

b. (15 points) What is Var(Y)? Give a numeric answer.

7. (20 points) Consider a hash table with n buckets. Assume that all buckets of the hash table are equally likely to get hashed to. Let X = the number of strings that are hashed up to and <u>including</u> the first time a string is hashed to a non-empty bucket (i.e., a "collision" occurs).

   a. (6 points) Say the number of buckets in the hash table $n = 4$. Calculate P(X = k) for $1 \leq k \leq 5$. In other words, calculate the probability that <u>exactly</u> k strings are hashed at the time that the first string is hashed to a non-empty bucket.

      P(X = 1) =

      P(X = 2) =

      P(X = 3) =

      P(X = 4) =

      P(X = 5) =

   b. (7 points) Still considering the case where $n = 4$, calculate E[X].

   c. (7 points) Now considering the general case of a hash table with n buckets, give a well-defined mathematical expression to compute E[X] in terms of n (for arbitrary positivex integer values of n). Of course, it's fine for your expression to include summations, products, etc. as long as it is well-defined.

8. (15 points) Consider a sample of independently and identically distributed (I.I.D.) random variables $X_1, X_2, \ldots, X_n$, that each have Geometric distributions. In other words, $X_i \sim \text{Geo}(p)$ for all $1 \leq i \leq n$.

   a. (11 points) Derive the Maximum Likelihood Estimate for the parameter p of the Geometric distribution. Give a numeric answer. Explicitly show all the steps in your derivation. (For this problem, the derivation is worth many more points than the final answer, so please be clear how you derived your final answer).

   b. (4 points) Say that we have a sample of five such I.I.D. Geometric variables with the following values: $X_1 = 4$, $X_2 = 3$, $X_3 = 4$, $X_4 = 2$, $X_5 = 7$. What value of p in the Geometric distribution would maximize the likelihood of these observations? Give a numeric answer.

9. (20 points) In this problem we will simultaneously estimate the difficulty of problem set questions and the skill level of each student. Consider a set of 200 students and 10 questions where each student answers each question. Let S_{ij} be an indicator variable which is 1 if student *i* answered question *j* correctly. You observe all S_{ij}.

   We are going to make the assumption that the probability (p_{ij}) that student *i* answers question *j* correctly is

   $$p_{i,j} = \sigma(a_i - d_j)$$

   where:

   σ is the sigmoid function,

   a_i is a parameter which represents a student's ability

   d_j is a parameter which represents a question's difficulty

   Use MLE to estimate the values for all parameters.

   a. Write the log likelihood for a single response S_{ij} in terms of p_{ij} (hint logistic regression also assumes that its output is a probability of a binary event)

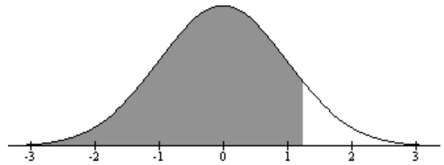
   b. What is the partial derivative of LL for a single response S_{ij} with respect to a_i?

   c. What is the partial derivative of LL for a single response S_{ij} with respect to *dj*?

   d. Explain briefly how you can estimate parameters given derivatives of log-likelihood with respect to those parameters.

Standard Normal Table

Note: An entry in the table is the area under the curve to the left of z, $P(Z \le z) = \Phi(z)$



Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7793	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8906	0.8925	0.8943	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998