ORCA Practice Review Questions

April 14, 2021

This is a set of review questions which are, I think, going to be helpful to guide your studying. It won't be an exact replica of the exam, nor will all of the review questions be in the style of questions on the exam, but I would say if you can answer these questions (or more accurately, understand the material that will allow to answer these questions), the exam will be very easy for you.

Part I: Data and Programming Ideas

1. What is the difference between long and wide data? What pandas functions are used to transform one data type to the other?

2. What is the difference between a pandas dataframe and a series?

3. What are the types of joins and how do they differ from one another?

4. What is a groupby and what does it do?

5. When I do arithmetic operations on two series with partial overlapping indexes, what is the index of the new series? What values are there for indexes that are in both series, and ones that are only in one.

Part II: Probability

1. Define a sample space. What are the three axioms of probability?

2. What does it mean when two events are independent?

3. What is the probability of the union of two events (not necessarily disjoint)?

4. What is conditional probability? What is Bayes theorem?

5. Suppose I roll two six-sided dice. X_1 is the result of the first die roll, X_2 is the result of the second die roll, and $X = X_1 + X_2$. Calculate the

following:

$$\begin{aligned} P(X = 5) \\ P(X_1 = 2 | X = 6) \\ P(X = 6 \cup X_2 = 3) \end{aligned}$$

6. What is a random variable? What is the difference between a discrete and continuous random variable?

7. Suppose a random variable X has the following density:

$$f(x) = \begin{cases} cx^2 & -1 \leq x \leq 1 \\ 0 & o/w \end{cases}$$

   Calculate c so that this is a proper density (integrates to 1), and calculate the expectation and variance of X.

8. Suppose a discrete random variable X has the following pmf:

$$p(x) = \begin{cases} 0.25 & x = 0 \\ 0.5 & x = 1 \\ 0.25 & x = 2 \end{cases}$$

   Calculate the expectation and variance of X.

9. If X and Y have the joint density:

$$f(x, y) = \begin{cases} x + y & x, y \in [0, 1] \\ 0 & o/w \end{cases}$$

   Calculate the covariance and correlation of X and Y.

Part III: Statistics

1. Suppose we have a sample $X_1, ..., X_n$ of random variable with the following density:

$$f(x) = \begin{cases} ax * \exp(-ax^2/2) & x > 0 \\ 0 & o/w \end{cases}$$

   Calculate the MLE for a.

2. What is a confidence interval? What is the difference between a one-sided lower/upper confidence interval, and a two-sided confidence interval?

3. What is a hypothesis test? Understand the following terms: critical region, level (of a test), p-value, and power

4. Suppose I have normally distributed data $X_1, ..., X_{10}$, with $\bar{X} = 1$ with unknown variance σ^2, but an estimate of $S^2 = 0.8$ from the data. How do I get a $(1-\alpha)100\%$ confidence interval (do not use the bootstrap) with $\alpha = 0.05$? What if I wanted to test:

$$H_0 : \mu = 0$$
$$H_A : \mu \neq 0$$

at $\alpha = .05$.

5. Describe the bootstrap procedure (write out an algorithm in pseudo-code).

Part IV: Machine Learning and Sklearn

1. How do ridge regression and LASSO differ from traditional linear regression? What is the difference between LASSO and ridge?

2. What kinds of problems do we use logistic regression for? What is the relationship between the covariates, regression coeffients, and the predicted probabilites in logistic regression?

3. Describe the process of cross-validation

4. Describe, at a high level, how a binary tree works? I.e. - given a datapoint, how do you end up with a prediction? What are the parts of a tree? Nodes, leaves, etc

5. What are boosting and bagging?

6. What is a random forest and how does it differ from a binary decision tree?

7. What are the primary metrics used to evaluate regression problems? What about classification problems?

8. Be prepared to write a simple pipeline potentially with a gridsearch using the tools from class (pipeline, columntransformer, polynomial features, one hot encoder, etc). It will not be that complicated!