Exam Questions:

**Question 1: Topics: Concept Drift/Bias-Variance/Experimental Design**

Let’s assume we have 3 time period T1-T3, ordered as follows:

Potential Training Data Available

**T1 (w/ TV Campaign)**

**T2 (No Campaign)**

**T3 (No Campaign)**

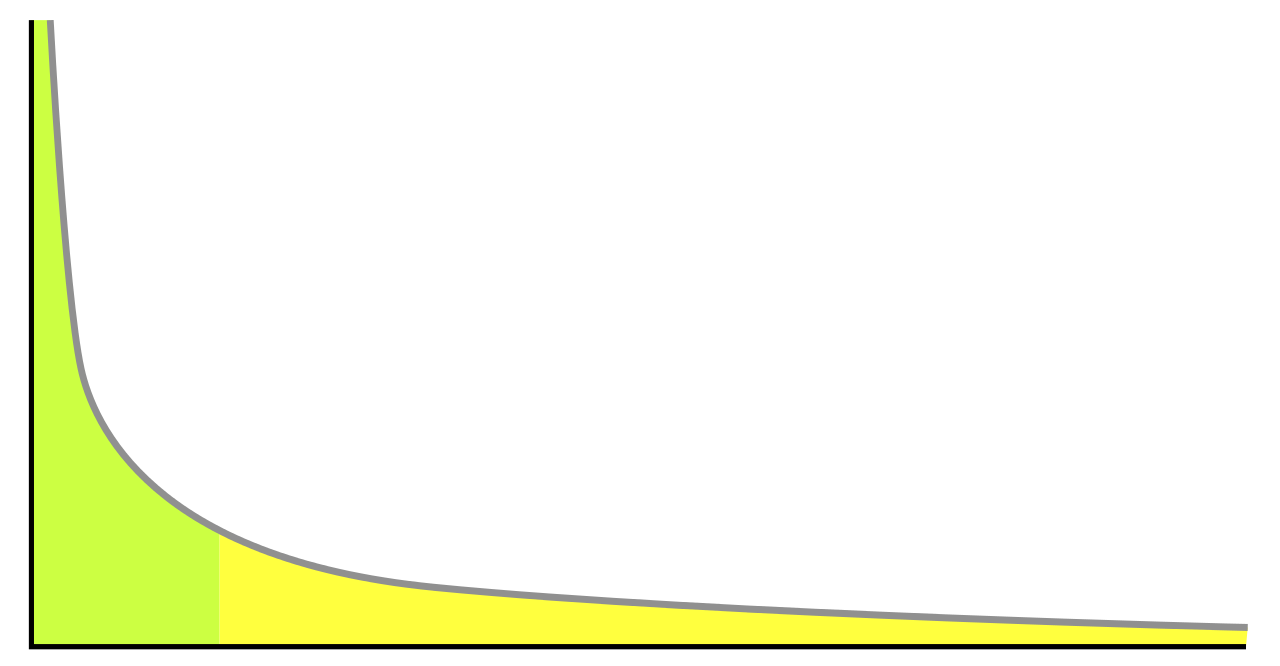
The end of T2 marks the present. You have been tasked to build a model that predicts P(Y|X) (don’t worry about what Y and X are right now), which will be applied in T3. You know from your marketing department that a major TV campaign ran for your company in T1 but not in T2. There is no plan to run one in T3. As the data scientist, you need to decide how much data to pull as a training set.

Knowing that the data from T1 and T2 differ because of the TV campaign, how will you decided what data to use? Propose a test/experimental design so that you can learn whether or not to use data from T1.

Explain the pros and cons of using just T2 or using both T1+T2 as your training set in terms of bias-variance tradeoffs.

**Question 2:**

Let’s assume your variable has the following distribution:



To be concrete, let’s assume the X-axis is a variable called total book sales and the Y-axis is the percentage of books with that value of sales. Your boss asks you to analyze this data and report back the average sales per book. Is the mean an appropriate distributional metric when the data is long-tailed as such? Why not? What other metrics could you propose in place of the mean?

**Question 3:**

Find the first and second derivative of the logistic log-likelihood (make this univariate)

**Question 4:**

List and describe 3 uses of the Singular Value Decomposition in data mining applications.

1. Your company invests in 10 different data sets to use as feature inputs in some sort of supervised classification model. As your data acquisition budget is finite, you need to evaluate whether each data source is worth the investment. As a data scientist, how would you:

a. Design an experiment to rank each data source according to the value it contributes.

b. How can you go from simply ranking the data sources to computing an ROI for each one?

2. On Regularization:

a. Explain the motivation behind regularization.

b. How do you determine the extent of regularization a model needs?

c. Beyond performance, why might you choose L1 vs. L2 regularization?

3. Your company wants to implement a classification model for making billions of predictions a day. The data you have for training has both millions of instances as well as millions of features. Additionally, predictions need to be made in 50ms or less.

a. What algorithms might be suitable for use in the production system?

b. Compare Random Forest and Logistic Regression in terms of:

i. training time

ii. scoring time

iii. model storage

c. How do you balance the trade-off between scalability and performance of the algorithm

4. On Model Evaluation:

a. When is it most appropriate to evaluate a model on the following measures: Area under the ROC curve, Accuracy and LogLoss/Cross Entropy?

b.  If you believe AUC is the most appropriate metric for a problem, why is it difficult to implement a training algorithm that optimizes AUC directly?

c. How would you explain AUC to a non data scientist?

