1. Consider the following semi-supervised variant of a support vector machine:

\[
\begin{align*}
\min_{w,b} & \quad \frac{1}{2} ||w||^2 \\
\text{s.t.} & \quad y_n(w^\top x_n + b) \geq 1 \quad (1 \leq n \leq L) \\
& \quad |w^\top x_n + b| \geq 1 \quad (L + 1 \leq n \leq L + U)
\end{align*}
\]  

where we have \( L \) labeled examples \((x_1, y_1), \ldots, (x_L, y_L)\) and \( U \) unlabeled examples \(x_{L+1}, \ldots, x_{L+U}\).

(a) Construct the Lagrange formulation (yes, again!) for this model. (10 points)

(b) In the Lagrange formulation, there is a term for the hinge loss on the labeled data points. There is also some term for a loss-like-thing on the unlabeled data points. What does this loss look like (i.e., draw a picture); it should be clear from this picture that this is non-convex. (10 points)

(c) (6350 only) What is the gradient of this formulation, as a function of \( w \) (don’t bother doing the case for \( b \))? If you were to run gradient descent, what (intuitively) would it do? (10 points)

2. Both semi-supervised learning and (pool-based) active learning make use of unlabeled data, but in different ways. Explain how unlabeled data helps in each of these learning paradigms? (10 points)

3. How are pool-based active learning and stream-based active learning different from each other? What are their pros and cons? In particular, compare them in terms of how well they can assess the informativeness of an example, and how efficient you expect their implementations to be?

In the class, we also looked at density based methods that try to deal with the issue of outlier examples which may seem to be the informative points but they aren’t actually. Do you think stream-based active learning can make use of density based methods? Why or why not? (20 points)

4. What are the things that a Naive Bayes classification model needs to estimate? How is it different from discriminative models such as logistic regression? Also explain the difference between the estimation equations in the case when the features in the data are discrete valued vs when they are continuous valued? In both cases, also explain the intuitive meaning of the estimation equations? (20 points)

5. Structured Prediction Basics:

(a) Structured Prediction algorithms such as MEMM, CRF, Structured-Perceptron/SVM, etc. use a feature function that depends on the outputs \( y \) as well (i.e. \( \phi(x, y) \)). What’s the reason for using such a feature function?

(b) Explain how the Maximum Entropy Markov Model (MEMM) is different from the Hidden Markov Model (HMM)? In particular, how is the label distribution estimated in both of these? Why is the MEMM more suitable than HMM in capturing feature dependencies? Which one has the larger number of parameters to be estimated (and why)? What are the single-label analogue models for HMM and MEMM?