CS 228T QUIZ 7

1. This question is about the pseudolikelihood objective.

(a) What is the pseudolikelihood objective? Why is the pseudolikelihood objective easier to evaluate than the likelihood objective?

(b) When does it coincide with the likelihood objective?

(c) When might we prefer to use the pseudolikelihood objective even when the likelihood objective is tractable to optimize? Give an example of a situation where pseudolikelihood provides a poor approximation to the likelihood objective.

2. This question is about structure learning in MRFs.

(a) Explain the score-based approach to structure learning in log-linear MRFs (*i.e.*, explain the setup of the problem we want to solve).

(b) What is the major drawback of using the likelihood score?

(c) What is the MAP score? Explain the differences between the use of ℓ_1 and ℓ_2 regularization.

(d) Explain the idea behind successor evaluation for optimizing the structure learning problem (for whatever objective function we have chosen). Discuss some differences between the use of this method for Markov random fields as opposed to Bayesian networks.

(e) Explain how the use of ℓ_1 regularization for structure learning can allow us to carry out structure learning via convex optimization rather than combinatorial search.

3. Both the grafting and the gain heuristics estimate the value of adding a feature to the log-linear model, but each uses a different heuristic approximation to obtain the estimate. Explain briefly what is the heuristic approximation used by each.

4. This question reviews some material about support vector machines.

(a) Consider the following formulation of the primal SVM. Let $\mathcal{D} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be a dataset, with $x_i \in \mathbb{R}^n$ and $y_i \in \{-1, 1\}$. The goal is to find a weight vector $w \in \mathbb{R}^n$ and offset $b \in \mathbb{R}$ such that

$$\mathbf{sign}(x_i^T w + b) = y_i$$

holds for many examples. Viewed as a function of x_i, the expression $x_i^T w + b$ is called a *discriminant function.* The condition that the sign of the discriminant function and the response should agree can be written as $u_i > 0$, where $u_i = y_i(x_i^T w + b)$ is the *margin* of the ith example.

In the theory of binary classification, loss functions φ are generally written as a function of the margin. A classification error is made if and only if the margin is negative, so φ should be positive and decreasing for negative arguments and zero or small for positive arguments. To find the parameters w and b, we minimize loss over the training set plus a regularization term on the weights. The primal SVM uses *hinge loss* $\varphi(u) = (1 - u)_+$, where $(z)_+ = \max(0, z)$, and

Date: September 29, 2020.

squared ℓ_2 regularization on the weights w:

$$\text{minimize} \quad \textstyle\sum_{i=1}^{m}(1 - y_i(x_i^T w + b))_+ + \lambda\|w\|_2^2,$$

where $\lambda > 0$ is a regularization parameter.

Explain how to rewrite this formulation to look like the formulation of the primal SVM problem in the reading. (Consider the transformation for piecewise-linear minimization or the epigraph transformation; these are described in Boyd and Vandenberghe.)

(b) Why can we restrict our attention to support vectors when classifying a new point with a dual SVM? In other words, explain why

$$w^T x + b = \sum_{s \in \mathcal{S}} \alpha_s y^{(s)} \langle x^{(s)}, x \rangle + b,$$

where $\mathcal{S}$ is the subset of the full training set consisting solely of the support vectors.

(c) What is a kernel function? Given an arbitrary function K, how could we check whether or not K is a valid kernel? What is the main benefit of being able to use kernel functions in support vector machines?

(d) Consider fitting a regularized SVM to a dataset that is linearly separable. Is the resulting decision boundary guaranteed to separate the classes?