COMS-4771 Spring 2008

Paper 3 Quiz Name: ______________________________

Taskar, Guestrin and Koller, *Max-Margin Markov Networks* (2003).

You have 30 minutes to complete the questions. The quiz is worth 10 points.

Question 1 (2 points): Give an example of a set of basis functions $f_k(\mathbf{x}, \mathbf{y})$ that you would use for the optical character recognition task (in equation (1) in the paper).

There are many possible answers. If the input $\mathbf{x}$ is a vector of pixel values and $\mathbf{y}$ is a word, the basis functions can be defined as in equation (2) in the paper, where a pairwise basis function $f_k(\mathbf{x}, y_i, y_j)$ is the indicator function

$$\mathbf{I}\left[\mathbf{x}_{p_1^k} = \text{on} \wedge \mathbf{x}_{p_2^k} = \text{on} \wedge y_i = c_1^k \wedge y_j = c_2^k\right],$$

where k ranges over $\langle p_1^k, p_2^k, c_1^k, c_2^k \rangle$ (pixel-character) combinations.

Question 2 (3 points):

- Give one reason to prefer a conditional model (a model of $p(\mathbf{y}|\mathbf{x})$) over a generative model (a model of $p(\mathbf{x}, \mathbf{y})$).
- Give an example of a regime where it makes sense to prefer a generative model over a conditional model.

A good approximation to the conditional distribution is certainly sufficient for classification. Given enough data, a conditional model will optimize the approximation for the optimal $p(\mathbf{y}|\mathbf{x})$, while the generative model may tune the approximation away from optimal conditional distribution, leading to worse discriminative performance.

However, a generative model typically needs fewer examples to find a good estimate of the joint distribution, so in a regime with very few training samples (relative to the number of parameters), generative models may outperform conditional models on the classification task.

Question 3 (3 points): The primal objective in the paper minimizes a convex upper bound on some loss function—what loss function? How can the loss be generalized without breaking the reduction to a polynomial size QP?

Hamming error. The loss can be generalizes to any decomposable loss without breaking the reduction.

Question 4 (2 points): Can a similar polynomial reduction be done for l_1 norm on w (instead of l_2)? Explain your answer.

Yes. A similar polynomial reduction can be done for any convex penalty including l_1.