CSE 5522 Homework 3

March 31, 2015

Problem 1 (2 points)

Consider the process of gradient-ascent training for a log-linear model with k features, given a dataset with N training instances. Assume for simplicity that the cost of computing a single feature over a single instance in our data set is constant, as is the cost of computing the expected value of each feature once we compute a marginal over all the variables in its scope. Assume it takes c time to compute all the marginals for each data case. Also assume that we need r iterations for the gradient process to converge.

(A) Using this notation, what is the time required to train an MRF in big-O notation?

(B) Using this notation, what is the time required to train a CRF in big-O notation?

Problem 2 (3 points)

Consider the log likelihood for a MRF in log-linear form:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_i \log P(y_i|\theta) = \frac{1}{N} \sum_i \left[\sum_c \theta_c^T \phi_c(y_i) - \log Z(\theta) \right]$$

Take the derivative and show that the gradient of the log likelihood is the difference between the expected feature vector according to the empirical distribution and the model's expectation of the feature vector:

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \left[\frac{1}{N} \sum_i \phi_c(y_i) \right] - \mathrm{E}[\phi_c(y)]$$

Problem 3 (3 points)

The Metropolis-Hastings algorithm is a member of the MCMC family; as such it is designed to generate samples x (eventually) according to target probabilities $\pi(x)$. (Typically we are interested in sampling from $\pi(x) = P(x|e)$.) Metropolis-Hastings operates in two stages. First it samples a new state x' from a *proposal distribution* $q(x'|x)$, given the current state x. Then, it probabilistically accepts or rejects x' according to the *acceptance probability*:

$$\alpha(x'|x) = \min\left(1, \frac{\pi(x')q(x|x')}{\pi(x)q(x'|x)}\right)$$

If the proposal is rejected, the state remains at x.

(A) Consider an ordinary Gibbs sampling step for a specific variable x_i. Show that this step, considered as a proposal is guaranteed to be accepted by Metropolis-Hastings. (Hence, Gibbs sampling is a special case of Metropolis-Hastings.)

(B) Show that the two-step process above, viewed as a transition probability distribution, is in detailed balance with π.

Problem 4 (3 points)

Recall the definition of *value of information*:

$$VPI_e(E_j) = \left(\sum_k P(E_j = e_{jk}|\mathbf{e})EU(\alpha_{e_{jk}}|\mathbf{e}, E_j = e_{jk})\right) - EU(\alpha|\mathbf{e})$$

(A) Prove that the value of information is nonnegative and order independent.

(B) Explain why it is that some people would prefer not to get some information–for example, not wanting to know the sex of their baby when an ultrasound is done.

(C) A function f on sets is **submodular** if, for any element x and sets A and B such that $A \subseteq B$, adding x to A gives a greater increase in f than adding x to B:

$$A \subseteq B \implies (f(A \cup x) - f(A)) \geq (f(B \cup x) - f(B))$$

Submodularity captures the intuitive notion of *diminishing returns*. Is the value of information, viewed as a function f on sets of possible observations, submodular? Prove this or find a counterexample.