67750: Advanced Practical Machine Learning	2019/20

Exercise 2 — Structured Prediction Theoretical Solutions

Dr. Omri Abend *TA: Omri Bloch*

1 Theoretical Questions

1.1 Energy Based Model Gradient

The MEMM, along with the MST parser, can be viewed as particular kinds of **energy based models** (EBM). These models are inspired by statistical mechanics and define an energy function, $E(x;\theta)$, such that the modeled distribution of the data follows the Boltzmann distribution:

$$p(x) = \frac{1}{Z}e^{-E(x;\theta)}$$

$$Z = \sum_x e^{-E(x;\theta)}$$

1. What is the energy function of the MEMM for a single transition, $E(y_t|y_{t-1}, x_{1:T};\theta)$?
2. Calculate the expected gradient of the log probability of a general energy based model and show that it is the difference between the gradient of the energy according to the model distribution and the true distribution. Formally, show that:

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\frac{\partial log(\mathbb{P}_\theta(x))}{\partial\theta}\Big] = \mathop{\mathbb{E}}_{x\sim\mathbb{P}_\theta}\Big[\frac{\partial E(x;\theta)}{\partial\theta}\Big] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\frac{\partial E(x;\theta)}{\partial\theta}\Big]$$

   In words, the gradient decreases the energy (increases the probability) of inputs which exist in the real distribution and increases the energy (decreases the probability) of inputs that exist in the model distribution.

Solution

1. $E(y_t|y_{t-1}, x_{1:T};\theta) = -w^T\phi(x_{1:T}, y_{t-1}, y_t, t)$
2. We first write the log probability according to the model:

$$log(\mathbb{P}_\theta(x)) = -E(x;\theta) - log(Z)$$

   Now, we take the gradient with respect to θ:

$$\frac{\partial log(\mathbb{P}_\theta(x))}{\partial\theta} = -\frac{\partial E(x;\theta)}{\partial\theta} - \frac{1}{Z}\sum_x e^{-E(x;\theta)}(-\frac{\partial E(x;\theta)}{\partial\theta}) = -\frac{\partial E(x;\theta)}{\partial\theta} + \sum_x \frac{1}{Z}e^{-E(x;\theta)}\frac{\partial E(x;\theta)}{\partial\theta} =$$

$$= -\frac{\partial E(x;\theta)}{\partial\theta} + \mathop{\mathbb{E}}_{x\sim\mathbb{P}_\theta}\Big[\frac{\partial E(x;\theta)}{\partial\theta}\Big]$$

So, taking the expectation over $x \sim \mathcal{D}$ (the actual distribution), we get:

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\frac{\partial log(\mathbb{P}_\theta(x))}{\partial\theta}\Big] = \mathop{\mathbb{E}}_{x\sim\mathbb{P}_\theta}\Big[\frac{\partial E(x;\theta)}{\partial\theta}\Big] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\Big[\frac{\partial E(x;\theta)}{\partial\theta}\Big]$$

Estimating this gradient for energy based models for learning is often challenging. In these cases it is often useful to approximate $\mathop{\mathbb{E}}_{x\sim\mathbb{P}_\theta}\Big[\frac{\partial E(x;\theta)}{\partial\theta}\Big]$ by different tricks, one of which is taking the argmax instead of the expectation.

1.2 MLE for HMM

In class, we saw the log likelihood function for the HMM:

$$\ell(S,\theta) = \sum_{i=1}^{N}\sum_{t=1}^{T^{(i)}} log(t_{y_{t-1}^{(i)},y_t^{(i)}}) + \sum_{i=1}^{N}\sum_{t=1}^{T^{(i)}} log(e_{y_t^{(i)},x_t^{(i)}})$$

Derive the MLE in class for $t_{i,j}$:

$$\hat{t}_{i,j} = \frac{\#(y_i \to y_j)}{\sum_k \#(y_i \to y_k)}$$

Solution

We can rewrite the log likelihood by summing over the possible transitions from y_{t-1} to y_t in the following way:

$$\ell(S,\theta) = \sum_i\sum_j \#(y_i \to y_j) log(t_{y_i,y_j}) + \sum_i\sum_{x_j} \#(y_i \to x_j) log(e_{y_i,x_j})$$

Now, we want to optimize for $t_{i,j}$, under the constraint that $\sum_k t_{i,k} = 1$. We can write the Lagrangian (ignoring the summation over the emission function since it will cancel out when we take the derivative):

$$\mathcal{L}(t,\lambda) = \sum_i\sum_j \#(y_i \to y_j) log(t_{i,j}) - \lambda(\sum_k t_{i,k} - 1)$$

Taking the derivative we find:

$$\frac{\partial\mathcal{L}}{\partial t_{i,j}} = \frac{\#(y_i \to y_j)}{t_{i,j}} - \lambda = 0 \to t_{i,j} = \frac{\#(y_i \to y_j)}{\lambda}$$

Finally, the λ which upholds our constraint is simply the sum of the possible transitions and we get:

$$\hat{t}_{i,j} = \frac{\#(y_i \to y_j)}{\sum_k \#(y_i \to y_k)}$$