netid: lw2350 | DS-GA 1008 Deep Learning
name: Lizhong Wang | October 31, 2020

Homework 3

Problem 1

(a) $nll = -\sum_{i=1}^{T} \log \tilde{y_i}[y_i]$, where $\tilde{y_i}[y_i]$ indicates the y_i^{th} element of $\tilde{y_i}$.

(b) Language model calculates the probability of a document by factorizing the joint probability into a product of conditional probability of each word conditioned on all its previous words. That is we were modeling $P(w_t|w_{1:t-1})$. Therefore, we can only go from left to right.

In the POS tagging task, we are just modeling the probability of the entire sequence $\boldsymbol{y}$ given the entire input sequence $\boldsymbol{x}$. Therefore, we can use information from both direction.

Problem 2

Pros

- Energy-based methods allows the model to make prediction under uncertainty / multimodality.
- Comparing to probabilistic models, EBM is not constrained by the axioms of the probability theory. Therefore, EBM gives more flexibility for choosing scoring function and objective function.

Cons

- Making inference with EBM can be slow because the inference step involves minimizing the energy function.

Problem 3

(a) $\check{y} = \arg\min_y E(x, y)$

We need to look up E $64 = 4^3$ times in order to obtain the best output sequence of POS tags because we have to check all possible combinations.

(b) (i) There are $\left(50000^{15}\, 20^{15}\right)$ different (x, y) pairs.

(ii) We need to look up E $\left(20^{15}\right)$ times in order to obtain the best output sequence of POS tags for a given sentence.

Problem 4

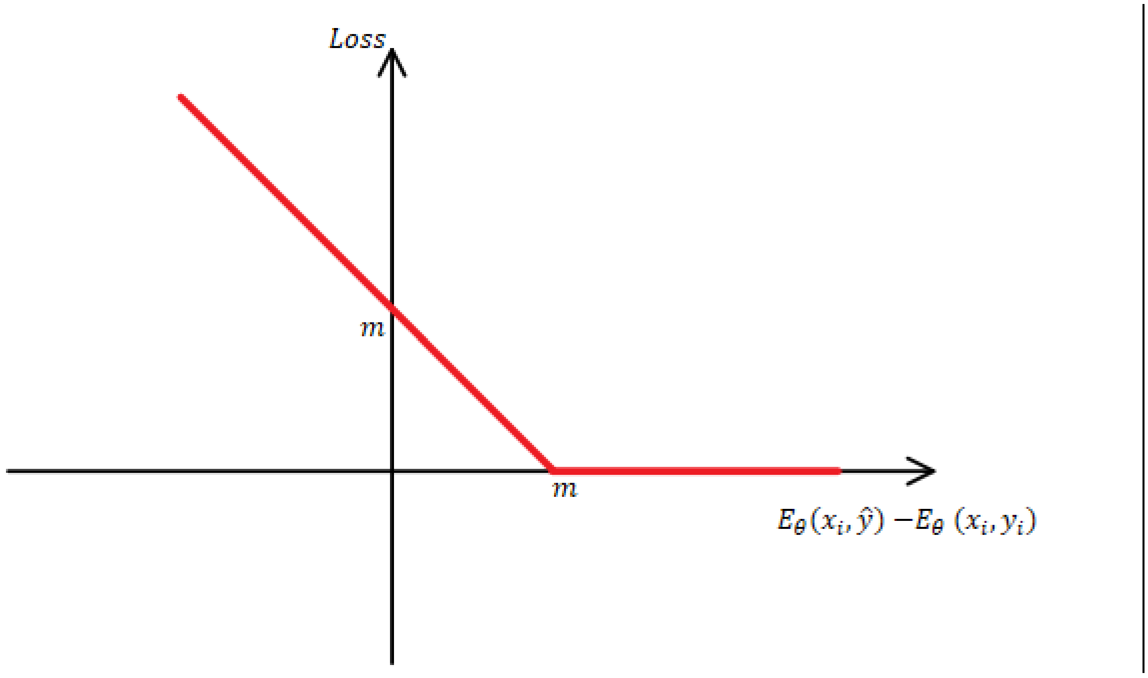
(a)

- $E_\theta(x_i, y_i)$ is the free energy of the gold-standard tag. In other words, $E_\theta(x_i, y_i)$ is the shortest distance from y_i, the gold-standard tag, to the model manifold.
- $\min_{y \neq y_i} E_\theta(x_i, y)$ is the smallest free energy of all the possible tags **excluding** y_i.

(b)

- According to a tutorial on EBM by Professor LeCun, we can call $\hat{y} = \arg\min_{y \neq y_i} E_\theta(x_i, y)$ the most offending incorrect tag.
- This is a loss function of contrastive method. For a given data point, we push the free energy of the gold-standard tag (y_i) down, while we pull the free energy of the most offending incorrect answer ($\hat{y}$) up.
- Ideally, for a given data point, we want to have the free energy of y_i to be the smallest among all possible tags. Also, we want the free energy of y_i (which is $E_\theta(x_i, y_i)$) to be smaller than the second smallest free energy (which is $\min_{y \neq y_i} E_\theta(x_i, y)$) by at least m. Minimizing this objective function will help us to achieve this goal as close as possible.
- The margin m allows us to control the level of contrast between $E_\theta(x_i, y_i)$ and $\min_{y \neq y_i} E_\theta(x_i, y)$.
- $[\cdot]_+$ makes the loss function not to penalize the model if the energy of the gold-standard tag is already smaller than the energy of the most offending tag by at least m units. Also $[\cdot]_+$ makes the loss function to penalize the model linearly.

(c) Plot:



- The loss is non-negative because of the $[\cdot]_+$.
- $Loss = 0 \ \rightarrow \ m + E_\theta(x_i, y_i) - E_\theta(x_i, \hat{y}) \leq 0 \ \rightarrow \ E_\theta(x_i, \hat{y}) - E_\theta(x_i, y_i) \geq m$
  Therefore, the loss equals to 0 when $E_\theta(x_i, \hat{y}) - E_\theta(x_i, y_i) \geq m$.

- After $E_\theta(x_i, \hat{y}) - E_\theta(x_i, y_i)$ becomes larger than m, then as $E_\theta(x_i, \hat{y}) - E_\theta(x_i, y_i)$ increases, the loss increase with a slop of 1.

(d) $y \leftarrow y - \eta \frac{\partial E_\theta(x,y)}{\partial y}$

(e) (i) We can use the edit distance between y and $y\prime$, which is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one vector into the other.
We can also use the square of the number of mismatches between y and $y\prime$, which will emphasize the large distance more.

(ii) Let the most offending tag be: $\hat{y} = \arg\min_{y \neq y_i} E_\theta(x_i, y)$.
Silimar to Objective (1), Objective (2) is also a contrastive loss function, and Objective (2) also contrast the gold-standard tag (y_i) and the most offending tag ($\hat{y}$).
However, instead of a constant m, Objective (2) uses a distance between y_i and $\hat{y}$, to dynamically control the level of contrast: the less similar/close y_i and $\hat{y}$ are, the larger.
In other words, the level of contrast between y_i and $\hat{y}$ is large when they are less similar, while the level of contrast is small when they are similar.

(iii) Objective (1) contrasts the data points equally with all other points, while Objective (2) contrast the data points more with the points that are less similar to the data points.
Suppose there are two non-data points A and B, and suppose A is much closer to the data manifold than B. Under Objective (1), there will be no loss incurred if both A and B are the most offending point for some data points and both A and B have energy m units smaller than the energy of their corresponding data points. However, this will not happen under Objective (2).
In short, Objective (2) takes the distance between y_i and $\hat{y}$ into consideration.
I think that with Objective (2), the slope of the energy function surface should be consistent with the distance between data pints and other points.

Problem 5

(a) Contrastive: Push down the energy of observed data points, while push up the energy of (all or part of) other points.

Architectural: Build a model such that the architecture of the model limits the volume of low energy space. For example, PCA and Gaussian Mixture Model.

Regularized: Use a regularization term to measure the volume the space that has low energy, and add this term to the objective function. For example, sparse coding.

(b) $F(y) = \min_z E(y,z) = \min_z [\ C(y, Dec(z)) + \lambda|z|_{L1}\ ]$

- The free energy of y, denoted as $F(y)$, is the minimum of the energy function of y and z over all possible z.
- The energy function, $E(y,z)$, has two terms.
- The first term is a cost function between y and the predicted y for a given z. The predicted y is denoted as $Dec(z)$.
- The second $\lambda|z|_{L1}$ regularize the information capacity of z. $|z|_{L1}$ is the L1 norm, and λ is the regularization coefficient.

(c) If we remove the regularize, then the model will overfit terribly. We can almost always perfectly reconstruct the original image, but this is due to memorizing but not learning. The energy function will have a mostly flat surface, which will not be useful.

Take the extreme case as an example. If z is the same or higher dimensional than y and the Decoder can implement identity function, then the decoder will become an expensive identity function, and we can always perfectly reconstruct the image. The surface of the energy function will be completely flat, and the energy function will be 0 for all y.

(d)
- The regularizer limits the information capacity of z. In other words, the regularizer limits the amount of configurations z can take, so that the model is forced to learn but not memorize.
- The regularizer makes the energy function have a bounded low energy space.
- Images that are similar to the training sampels will have small energy, while images that are dissimilar to the training images will have larger energy.

(e) In k-means, a point has to belong to one and only one of the k possible classes. Therefore, the latent z is a k-dimensional one-hot vector. The model design bounded the information capacity of z: for each y, the latent z can only take one of k possible states. Therefore, k-means is an architectural method. The free energy function is:

$$F(y) = \min_z ||y - wz||^2$$

Assume a data point is d-dimensional and we have k classes, then:

$$y \in \mathbb{R}^d, \quad z \text{ is a one-hot vector and } z \in \mathbb{R}^k, \quad w \text{ is a matrix and } w \in \mathbb{R}^{d\times k}$$