Quiz-2

Sunday 23$^{\text{rd}}$ October, 2016

(**15 Marks**). Please answer **to the point** in the limited space provided for each question. You can do rough work in a separate sheet of paper.

Problem 1. Logistic Regression and Maximum Entropy Classifier
The Logistic Regression classifier also goes under another name called the Maximum Entropy Classifier where the goal is to prefer the most uniform models that also satisfy any given constraints. More specifically, the goal in Maximum Entropy Classification is to find the probability distribution $\Pr\left(Y = c|\phi\left(\mathbf{x}\right)\right)$ for $c = [1...K]$ that maximizes the entropy

$$E\left(\Pr(.)\right) = -\left[\frac{1}{m}\sum_{c=1}^{K}\sum_{i=1}^{m}\Pr\left(Y = c\middle|\phi\left(\mathbf{x}^{(i)}\right)\right)\log\Pr\left(Y = c\middle|\phi\left(\mathbf{x}^{(i)}\right)\right)\right] \tag{1}$$

such that, for every feature ϕ_j for $j = [1...n]$ and every class $c = [1..K]$

$$\sum_{i=1}^{m}\phi_j\left(\mathbf{x}^{(i)}\right)\Pr\left(Y = c\middle|\phi\left(\mathbf{x}^{(i)}\right)\right) = \sum_{i=1}^{m}\phi_j\left(\mathbf{x}^{(i)}\right)\delta(y^{(i)}, c) \tag{2}$$

where $\delta(y^{(i)}, c) = 1$ if and only if $y^{(i)} = c$ and $\delta(y^{(i)}, c) = 0$ otherwise.

Now answer the following questions

1. Interpret the optimization problem (1) and the constraint (2) in plain English words while clearly stating the intuition.
2. Prove that the probability distribution $\Pr\left(Y = c|\phi\left(\mathbf{x}\right)\right)$ of the solution that maximizes the entropy in (1) subject to the constraint set (2) turns out to have the form of logistic regression[1].
3. How can one introduce regularization into the entropy objective (1)? What will be the result of doing so on the form of the resulting probability distribution $\Pr\left(Y = c|\phi\left(\mathbf{x}\right)\right)$ at optimality? (**10 Marks**)

[1] Optional point of reference: Recall Solution to Problem 3 of Tutorial 7, where we restated the problem of finding optimal solution $\mathbf{w}$ to the regularized cross entropy as that of finding a function from a function space that is smooth enough and minimizes the objective. In this problem, we are similarly seeking a characterization of family of probability distributions that maximize the entropy.

Solution:

1. The main idea behind maximum entropy is that one should prefer the most uniform models that also satisfy the data driven constraints which are captured in (2). For example, consider a four-way text classification task where we are told only that on an average 60% of documents with the word “advisor” in them are in the *student* class. Thus, given a document with “advisor” in it, we would say it has a 60% chance of being a *student* document, and a 40% chance for each of the other three classes. If a document does not have“advisor” we would guess the uniform class distribution, that is, 25% each as per (1). This is exactly the maximum entropy model (1) that conforms to our known constraint (2)

2. The key is to understand that the optimization problem in (1) is for the set of variables p_{ci} where $p_{ci} = \Pr\left(Y = c|\phi\left(\mathbf{x}^{(i)}\right)\right)$. Simplfying (1) and (2) in terms of these variables

$$E\left(\Pr(.)\right) = -\left[\frac{1}{m}\sum_{c=1}^{K}\sum_{i=1}^{m} p_{ci}\log p_{ci}\right] \tag{3}$$

such that, for every feature ϕ_j for $j = [1...n]$ and every class $c = [1..K]$

$$\sum_{i=1}^{m}\phi_j\left(\mathbf{x}^{(i)}\right)p_{ci} = \sum_{i=1}^{m}\phi_j\left(\mathbf{x}^{(i)}\right)\delta(y^{(i)}, c) \tag{4}$$

and for each $i = [1..m]$

$$\sum_{c=1}^{K} p_{ci} = 1 \tag{5}$$

Let w_{cj} be the lagrange multiplier corresponding to (4) for a specific value of c and j and λ be the lagrange multiplier corresponding to (5). We know that a necessary condition for optimality of (3) subject to (4) and (5) is that the gradient (or every partial derivative) of the Langrange with respect to all the p_{ci} should be 0. That is,

$$-\log p_{ci} - 1 + \sum_{j=1}^{m} w_{cj}\phi_j\left(\mathbf{x}^{(i)}\right) + \lambda = 0$$

that is,

$$\log p_{ci} = \lambda - 1 + \sum_{j=1}^{m} w_{cj}\phi_j\left(\mathbf{x}^{(i)}\right)$$

That is,

$$p_{ci} = \exp\left(\lambda - 1 + \sum_{j=1}^{m} w_{cj}\phi_j\left(\mathbf{x}^{(i)}\right)\right) = \exp\left(\lambda - 1\right) \times \exp\left(\sum_{j=1}^{m} w_{cj}\phi_j\left(\mathbf{x}^{(i)}\right)\right)$$

Now from (5),

$$\sum_{k=1}^{K} p_{ki} = \sum_{k=1}^{K} \exp\left(\lambda - 1\right) \times \exp\left(\sum_{j=1}^{m} w_{kj}\phi_j\left(\mathbf{x}^{(i)}\right)\right) = \exp\left(\lambda - 1\right) \times \sum_{k=1}^{K} \exp\left(\sum_{j=1}^{m} w_{kj}\phi_j\left(\mathbf{x}^{(i)}\right)\right) = 1$$

Therefore

$$\exp\left(\lambda - 1\right) = \frac{1}{\sum_{k=1}^{K} \exp\left(\sum_{j=1}^{m} w_{kj}\phi_j\left(\mathbf{x}^{(i)}\right)\right)}$$

Thus,

$$\Pr\left(Y = c|\phi\left(\mathbf{x}^{(i)}\right)\right) = p_{ci} = \frac{\exp\left(\sum_{j=1}^{m} w_{cj}\phi_j\left(\mathbf{x}^{(i)}\right)\right)}{\sum_{k=1}^{K} \exp\left(\sum_{j=1}^{m} w_{kj}\phi_j\left(\mathbf{x}^{(i)}\right)\right)}$$

Also, $\sum_{j=1}^{m} w_{cj}\phi_j\left(\mathbf{x}^{(i)}\right) = \mathbf{w}_c^T\phi\left(\mathbf{x}^{(i)}\right)$. Thus

$$\Pr\left(Y = c|\phi\left(\mathbf{x}^{(i)}\right)\right) = p_{ci} = \frac{\exp\left(\mathbf{w}_c^T\phi\left(\mathbf{x}^{(i)}\right)\right)}{\sum_{k=1}^{K} \exp\left(\mathbf{w}_k^T\phi\left(\mathbf{x}^{(i)}\right)\right)}$$

which is exactly of the form of Logistic regression!

3. Any one of the two options could serve the purpose.

   - Note that maximum entropy without any constraints results when the distribution is uniform. Thus, maximizing entropy is the same as minimizing the (probabilistic) distance of the target desired distribution from a uniform prior. One could therefore introduce regularization into the entropy maximization in (1) by changing the objective from maximizing entropy (that is minimizing the distance of the target distribution from the uniform distribution) to **minimizing the distance of the target distribution from some prior distribution** (such as Gaussian), subject to the constrains in (2). Specifically, such one such distance is called the **KL divergence**.
   - One could introduce regularization into the entropy maximization via the constraints (2) by adding a small factor η to each constraint on the right hand side. This amounts to the assumption that by default certain additional number of observations η for each feature need to be accounted for, even if the number of such observations in the data are fewer. This allows for sparse features to be accounted for.

Problem 2. Answer the following questions on neural networks:

1. Can you think of a smaller and more compact network for XOR than was discussed in class? What is the smallest possible network?

   Solution: You will need atleast two perceptrons in the neural network, since XOR is not linearly seperable. It turns out that you can indeed construct a XOR with 2 perceptrons as follows

2. Now how about designing an N-way XOR using neural networks? Recall that an $N-$way XOR outputs a 1 if an only if an odd number of inputs have 1.

   Solution: We can recursively partition each problem into two subproblems with half the size as follows:

   $XOR(x_1...x_n) = XOR(XOR(x_1...x_{n/2}), XOR(x_{n/2+1}, x_n)$

   Assuming that $XOR(a, b)$ can be represented using two perceptrons, it will require $\log N$ such units (that is $2 \log N$ perceptrons) to represent $XOR(x_1...x_N)$

(**5 Marks**)