Logistic Regression, Expectation Maximization and Probability Bounds

Submit a PDF of your answers to Canvas.

1. In this problem we will train a binary logistic regression classifier. Download the starter notebook and the associated dataset.

Training in binary logistic regression involves minimizing the following:

$$\min_{\boldsymbol{w}} \sum_i \log\left(1 + e^{-y_i \boldsymbol{x}_i^T \boldsymbol{w}}\right)$$

where $y_i \in \{-1, 1\}$ is the class label, $\boldsymbol{x}_i$ is the feature vector, $\boldsymbol{w}$ is the unknown weight vector, and the sum is taken over the training data. The gradient of this expression with respect to $\boldsymbol{w}$ is

$$\nabla_w \ell(\boldsymbol{w}) = \sum_i \frac{-y_i}{1 + e^{y_i \boldsymbol{w}^T \boldsymbol{x}_i}} \boldsymbol{x}_i.$$

a) Write a function to compute the logistic loss for a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Your function should take a weight vector $\boldsymbol{w}$ as input and return a scalar.

b) Write a function to compute the gradient of the logistic loss evaluated at $\boldsymbol{w}$. Your function should a weight vector $\boldsymbol{w}$ as input and return a column vector.

c) Run gradient descent on the provided dataset using the functions you created. Make sure to specify an appropriate stopping condition.

d) What is the error rate on the (training) dataset?

e) **Optional.** Improve the gradient descent algorithm by using Newton's method. Newton's works by estimating the function using a second order Taylor series approximation. You will need to compute the Hessian of the logistic loss function.

2. *K means and EM.* The K means algorithm is an example of an expectation maximization (EM) algorithm. K means involves finding the closest cluster center for each each data point as:

$$z_i = \arg\min_k ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||_2^2$$

Note that z_i indicates the cluster to which point i is assigned.

The EM algorithm can also be used for clustering. In Gaussian mixture models, the *responsibility* of each cluster k for a data point i is computed as

$$r_{ik} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_{k'}, \Sigma_{k'})}.$$

This is often referred to as *soft* assignment, as each cluster (each mixture component) takes some responsibility for each data point. Conversely, hard assignment involves assigning each data point to the cluster that has the largest *responsibility*:

$$z_i = \arg\max_k r_{ik}.$$

a) Show that the two expressions for z_i are equivalent when $\boldsymbol{\Sigma}_k = \boldsymbol{I}$ and $\pi_k = 1/K$.

b) Argue that, in some cases, the soft and hard assignments are nearly equivalent:

$$r_{ik} \approx \begin{cases} 1 & \text{if } \pi_k \mathcal{N}(\boldsymbol{x}_i|\mu_k, \Sigma_k) \gg \pi_{k'} \mathcal{N}(\boldsymbol{x}_i|\mu_{k'}, \Sigma_{k'}) \qquad k' \neq k \\ 0 & \text{else.} \end{cases}$$

3. *Empirical CDFs.* Recall that the cumulative distribution function of a random variable with pdf $f(x)$ is given by $F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(y)dy$.

a) Use the definition of expectation to show that $F(x) = E[\mathbb{I}\{X \leq x\}]$.

b) Let $X_1, X_2, \ldots, X_n$ denote i.i.d samples from $f(x)$. The empirical distribution of continuous data is usually defined as $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}$. What is $\mathbb{E}[\widehat{F}_n(x)]$?

c) Fix an x, and we have that $\text{var}(\widehat{F}_n(x)) = \mathbb{E}[(\widehat{F}_n(x) - F(x))^2]$. Show that $\text{var}(\widehat{F}_n(x)) = \frac{F(x)(1-F(x)}{n}$.

d) Show that for all x, $\mathbb{E}[(\widehat{F}_n(x) - F(x))^2] \leq \frac{1}{4n}$.