Math 7243 Machine Learning - Homework 3

For programming questions, you can only use numpy library.

Question 1. Softmax regression Recall the setup of logistic regression: We assume that the posterior probability is of the form

$$p(Y = 1|\vec{x}) = \frac{1}{1 + e^{-\beta^T \vec{x}}}$$

This assumes that $Y|X$ is a Bernoulli random variable. We now turn to the case where $Y|X$ is a multinomial random variable over K outcomes. This is called softmax regression, because the posterior probability is of the form

$$p(Y = k|\vec{x}) = \frac{e^{\beta_k^T \vec{x}}}{\sum_{j=1}^K e^{\beta_j^T \vec{x}}}$$

which is called the softmax function. Assume we have observed data $D = \{\vec{x}^{(i)}, y^{(i)}\}_{i=1}^N$. Our goal is to learn the weight $\beta_1, ..., \beta_K$.

(1) Find the negative log likelihood of the data $l(\beta_1, ..., \beta_K) = -\log L(\beta_1, ..., \beta_K) = -\log P(Y|X)$

$$-\log \mathbb{P}(Y|X) = -\log \prod_{i=1}^N \mathbb{P}(y_i|x_i) = -\log \prod_{i=1}^N \prod_{k=1}^K \left(\frac{e^{\beta_k^T x_i}}{\sum_{j=1}^K e^{\beta_j^T x_i}} \right)^{1\{y_i=k\}}$$

$$= -\sum_{i=1}^N \sum_{k=1}^K 1\{y_i = k\} \left(\beta_k^T x_i - \log \left(\sum_{j=1}^K e^{\beta_j^T x_i} \right) \right)$$

$$= -\sum_{i=1}^N \sum_{k=1}^K 1\{y_i = k\} \beta_k^T x_i + \sum_{i=1}^N \log \left(\sum_{j=1}^K e^{\beta_j^T x_i} \right)$$

(2) We want to minimize the negative log likelihood. To combat overfitting, we put a regularizer on the objective function. Find the **gradient** w.r.t. β_k of the regularized objective

$$l(\beta_1, ..., \beta_K) + \lambda \sum_{k=1}^K ||\beta_k||^2$$

$$\nabla_{\beta_k} - \log \mathbb{P}(Y|X) = 2\lambda\beta_k - \sum_{i=1}^N 1\{y_i = k\} x_i + \sum_{i=1}^N \frac{e^{\beta_k^T x_i}}{\sum_{j=1}^K e^{\beta_j^T x_i}} x_i$$

Note that we can use the definition of $\mu_k(x_i)$ here to save a bunch of writing.

$$= 2\lambda\beta_k + \sum_{i=1}^N (\mu_k(x_i) - 1\{y_i = k\}) x_i$$

(3) State the gradient updates for both batch gradient descent and stochastic gradient descent.

Batch gradient descent:

$$\beta_k^{(t+1)} = \beta_k^{(t)} - \eta \left(2\lambda\beta_k^{(t)} + \sum_{i=1}^{N} \left(\mu_k(x_i) - 1\{y_i = k\}\right) x_i \right)$$

Stochastic gradient descent:

$$\beta_k^{(t+1)} = \beta_k^{(t)} - \eta \left(2\lambda\beta_k^{(t)} + \left(\mu_k(x_i) - 1\{y_i = k\}\right) x_i \right)$$

Question 2. - Linear Discriminant Analysis: Consider the categorical learning problem consisting of a data set with two labels:

Label 1:

X_1	3.81	0.23	3.05	0.68	2.67
X_2	-0.55	3.37	3.53	1.84	2.74

Label 2:

X_1	-2.04	-0.72	-2.46	-3.51	-2.05
X_2	-1.25	-3.35	-1.31	0.13	-2.82

a) For each label above, the data follow a multivariate normal distribution Normal(μ_i, Σ) where the covariance Σ is the same for both label 1 and for label 2. Fit a pair of Guassian discriminant functions to the labels by computing the covariances, means, and proportions of datapoints as discussed in the Linear Discriminant Analysis section. You may use a computer, but you should **not** use an LDA solver. You should report the values for μ_i and Σ.

$\mu_1 = \begin{pmatrix} 2.088 \\ 2.186 \end{pmatrix}, \mu_2 = \begin{pmatrix} -2.156 \\ -1.72 \end{pmatrix}$.

$\Sigma = \frac{1}{10-2} \sum_{i=1}^{10} (X^{(i)} - \mu_k)(X^{(i)} - \mu_k)^T = \begin{pmatrix} 1.709575 & -1.23013 \\ -1.23013 & 2.349865 \end{pmatrix}$.

$\phi_1 = \phi_2 = \frac{5}{10} = \frac{1}{2}$.

$P(X|\text{Label} = 1) = \frac{1}{(2n)^{n/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1} (x - \mu_1))$
$P(X|\text{Label} = 2) = \frac{1}{(2n)^{n/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(x - \mu_2)^T \Sigma^{-1} (x - \mu_2))$

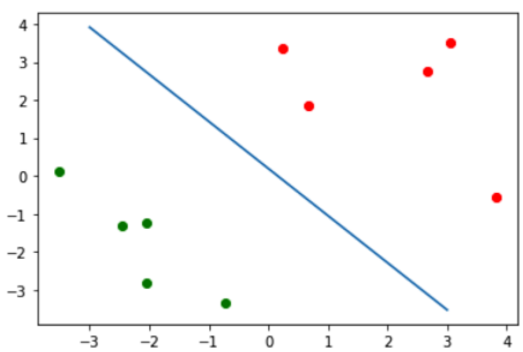
b) Give the **formula for the line** forming the discretion boundary.

The line forming the discretion boundary is $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ such that $\log \frac{P(\text{Label}=2|X)}{P(\text{Label}=1|X)} = 0$.

$P(\text{Label} = k|X) = \frac{P(X|\text{Label}=k)P(\text{Label}=k)}{P(X)}$.

$\log P(\text{Label} = k|X) = \log P(X|\text{Label} = k) + \log P(\text{Label} = k) - \log P(X)$
$= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(X - \mu_k)^T\Sigma^{-1}(X - \mu_k) + \log\phi_k + \text{ constant}$
$= X^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\phi_k - \frac{1}{2}\log|\Sigma| + \text{ constant}$

Hence, $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ such that $\log P(\text{Label} = 1|X) = \log P(\text{Label} = 2|X)$.
$X^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 = X^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2$
$\begin{pmatrix} x_1 & x_2 \end{pmatrix}\begin{pmatrix} 3.03331817 \\ 2.51817686 \end{pmatrix} - 5.91915147956369 = \begin{pmatrix} x_1 & x_2 \end{pmatrix}\begin{pmatrix} -2.86820579 \\ -2.23343298 \end{pmatrix} - 5.012678200684864$
$\begin{pmatrix} x_1 & x_2 \end{pmatrix}\begin{pmatrix} 3.03331817 + 2.86820579 \\ 2.51817686 + 2.23343298 \end{pmatrix} = -5.012678200684864 + 5.91915147956369$
$5.90152396x_1 + 4.75160984x_2 = 0.9064732788788268$
$x_2 = \frac{0.9064732788788268 - 5.90152396x_1}{4.75160984}$



c) (bonus) Try the **QDA** method for this question and obtain an quadratic boundary.

Question 3. later

Question 4. later