Homework 2
Due: Midnight, Mar 15st, 2022

Theory

1. Assume X are observations and Z are hidden variables. $p(X)$ and $p(Z)$ are the probability distributions over X and Z. We know:

$$p(Z|X) = \frac{p(X|Z)p(Z)}{p(X)} = \frac{p(X|Z)p(Z)}{\int_Z p(X,Z)}$$

We try to use $q(Z)$ to approximate $p(Z|X)$. Please prove:

$$KL(q(Z)||p(Z|X)) = -E_q[\log p(X,Z)] + E_q[\log q(Z)] + \log p(X),$$

where $KL(\cdot)$ is the Kullback-Leibler (KL) divergence. We usually define $L = E_q[\log p(X,Z)] - E_q[\log q(Z)]$ as the variational lower bound. (Hint: https://xyang35.github.io/2017/04/14/variational-lower-bound/)

2. Let $\delta(x)$ be a Bayes classifier for binary classification:

$$\delta(x) = \begin{cases} +1, \text{if } P(Y=+1 \mid X=x) \geq P(Y=-1 \mid X=x) \\ -1, \qquad\qquad otherwise. \end{cases}$$

Let $R(t) = P(Y \neq t(x)) = \mathbb{E}X[P(Y \neq t(x) \mid X=x)]$ be the error rate for a classifier $t(x)$. Please prove $R(\delta) \geq R(f)$, where $f(x)$ is any binary classifier. (Hint: https://mlweb.loria.fr/book/en/bayesclassifier.html)

3. Let $\{(x_i, y_i)\}_{i=1}^n$ be a set of n paired samples, where $x_i \in \mathbb{R}^d$ is the feature vector of the ith sample and y_i is its label. In the class, we know the log likelihood function of the logistic regression is $\ell(\beta) = \sum_{i=1}^n y_i \beta^T x_i - \log(1 + e^{\beta^T x_i})$. Please compute $\frac{\partial \ell(\beta)}{\partial \beta}$ and $\frac{\partial \ell(\beta)}{\partial \beta \, \partial \beta^T}$. (Hint: https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf)

Programing

1. Analysis scATAC-seq data.

1.1 Download the single-cell ATAC-seq data from the GEO website:
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126074
You need to download the following three files

GSE126074_AdBrainCortex_SNAREseq_chromatin.barcodes.tsv.gz	54.1 Kb	(ftp)(http)	TSV
GSE126074_AdBrainCortex_SNAREseq_chromatin.counts.mtx.gz	77.7 Mb	(ftp)(http)	MTX
GSE126074_AdBrainCortex_SNAREseq_chromatin.peaks.tsv.gz	1.9 Mb	(ftp)(http)	TSV

1.2 Use episcanpy to analyze the data. Follow this Tutorial: https://nbviewer.org/github/colomemaria/epiScanpy/blob/master/docs/tutorials/Buenrostro_PBMC_data_processing.html to find and visualize the clusters of this scATAC-seq data.

1.3 Use cistopic to do dimension reduction of the same data. Then find and visualize the clusters of this scATAC-seq data. Compare with the result obtained in 1.2.

2. Benchmarking different classification method on cancer type prediction.
2.1 Download data from cBioportal (https://www.cbioportal.org/datasets). Download the following data:

Breast Invasive Carcinoma (TCGA, Cell 2015)	TCGA, Cell 2015	818	817	816	817

Extract patients with Breast Invasive Ductal Carcinoma and Breast Invasive Lobular Carcinoma. Obtain their mutation data.

2.2 Use five-fold cross-validation to compare LDA, logistic regression, and Naive Bayesian in terms of F1 score.