Introduction to Machine Learning **Fall Semester, 2023/24**

Homework 1: January 3, 2024

Due: January 24, 2023

Linear Algebra

1. (**15 pts**) A symmetric matrix A over $\mathbb{R}$ is called *positive semidefinite* (PSD) if for every vector $\mathbf{v}$, $\mathbf{v}^\mathsf{T} A\mathbf{v} \geq 0$.

   (a) Show that a symmetric matrix A is PSD if and only if it can be written as $A = XX^T$, if and only if all of its eigenvalues are non-negative.
   Hint: Recall that a real symmetric matrix A can be decomposed as $A = QDQ^T$, where Q is an orthogonal matrix whose columns are eigenvectors of A and D is a diagonal matrix with eigenvalues of A as its diagonal elements.

   (b) Show that for all $\alpha, \beta \geq 0$ and PSD matrices $A, B \in \mathbb{R}^{n\times n}$, the matrix $\alpha A + \beta B$ is also PSD. Does this mean that the set of all $n \times n$ PSD matrices over $\mathbb{R}$ is a vector space over $\mathbb{R}$?

Calculus and Probability

1. (**15 pts**) Let $X_1, ..., X_n$ be i.i.d $U([0, 1])$ (uniform) continuous random variables. Let $Y = \max(X_1, ..., X_n)$.

   (a) What is the PDF of Y? Write the mathematical formula and plot the PDF as well. Compute $\mathbb{E}[Y]$ and $\mathrm{Var}[Y]$ - how do they behave as a function of n as n grows large?

   (b) (**No need to submit**) Verify your answer empirically using Python.

Optimal Classifiers and Decision Rules

1. (**15 pts**)

   (a) Let X and Y be random variables where Y can take values in $\mathcal{Y} = \{1, \ldots, L\}$. Let ℓ_{0-1} be the 0-1 loss function defined in class. Show that $h = \arg\min_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}\left[\ell_{0-1}(Y, f(X))\right]$ is given by
   $$h(x) = \arg\max_{i\in\mathcal{Y}} \mathbb{P}[Y = i|X = x]$$
   (In class you proved this for binary classification, i.e. $\mathcal{Y} = \{0, 1\}$, here you are asked to generalize the result to L labels.)

   (b) Let X and Y be random variables where Y can take values in $\mathcal{Y} = \{0, 1\}$. Let Δ be the following asymmetric loss function:
   $$\Delta(y, \hat{y}) = \begin{cases} 0 & y = \hat{y} \\ a & y = 0, \hat{y} = 1 \\ b & y = 1, \hat{y} = 0, \end{cases}$$

where $a, b \in (0, 1]$ (note that this loss function generalizes the 0-1 loss defined in class). Compute the optimal decision rule h for the loss function Δ, i.e. the decision rule which satisfies:

$$h = \arg \min_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}\left[\Delta(Y, f(X))\right]$$

2. (**15 pts**) Let X and Y be random variables where X can take values in some set $\mathcal{X}$ and Y can take values in $\mathcal{Y} = \{0, 1\}$ (i.e. binary label space). Assume we wish to find a predictor $h : \mathcal{X} \to [0, 1]$ (note that the hypothesis can output any number between 0 and 1) which minimizes $\mathbb{E}[\Delta_{log}(Y, h(X))]$, where Δ_{log} is the following loss function known as the *log-loss*:

$$\Delta_{log}(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}).$$

Find the predictor $h : \mathcal{X} \to [0, 1]$ which minimizes $\mathbb{E}[\Delta_{log}(Y, h(X))]$ (the final answer should be very intuitive).

Note: This loss function may seem odd at first, but it is very important and we'll discuss it further in the future.

3. (**10 pts**)

Let X and Y be random variables taking values in $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{0, 1\}$ respectively, and assume that given $Y = 0$, X is distributed normally with mean μ and variance σ_0^2, i.e. $X \sim \mathcal{N}(\mu, \sigma_0^2)$, and similarly, given $Y = 1$, $X \sim \mathcal{N}(\mu, \sigma_1^2)$, where $\sigma_0 \neq \sigma_1$. Also assume $\Pr[Y = 1] = p_1$.

Find the optimal decision rule for this distribution and the zero-one loss, i.e. find $h : \mathbb{R} \to \{0, 1\}$ which minimizes $\mathbb{E}[\ell_{0-1}(Y, h(X))]$ where ℓ_{0-1} is the zero-one loss defined in class (write the decision rule only in terms of $x, \mu, \sigma_0, \sigma_1$ and p_1).

Programming Assignment

1. **Visualizing the Hoeffding bound** (**10 pts**).

   (a) Use **numpy** to generate an $N \times n$ matrix of samples from $Bernoulli(1/2)$. Calculate for each row i the empirical mean, $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{i,j}$, where $N = 200000$ and $n = 20$.

   (b) Take 50 values of $\epsilon \in [0, 1]$ (**numpy.linspace(0,1, 50)**), and calculate the empirical probability that $|\bar{X}_i - 1/2| > \epsilon$. Plot the empirical probability as a function of ϵ.

   (c) Add to your plot the Hoeffding bound of that probability, as a function of ϵ.

   Submit your plots (no need to submit code for this question).

2. **Nearest Neighbor** (**20 pts**). In this question, we will study the performance of the Nearest Neighbor (NN) algorithm on the MNIST dataset. The MNIST dataset consists of images of handwritten digits, along with their labels. Each image has 28×28 pixels, where each pixel is in gray-scale, and can get an integer value from 0 to 255. Each label is a digit between 0 and 9. The dataset has 70,000 images. Although each image is square, we treat it as a vector of size 784.

   The MNIST dataset can be loaded with `sklearn` as follows:

```
>>> from sklearn.datasets import fetch_openml
>>> mnist = fetch_openml('mnist_784', as_frame=False)
>>> data = mnist['data']
>>> labels = mnist['target']
```

   Loading the dataset might take a while when first run, but will be immediate later. See `http://scikit-learn.org/stable/datasets.html` for more details. Define the training and test set of images as follows:

```
>>> import numpy.random
>>> idx = numpy.random.RandomState(0).choice(70000, 11000)
>>> train = data[idx[:10000], :].astype(int)
>>> train_labels = labels[idx[:10000]]
>>> test = data[idx[10000:], :].astype(int)
>>> test_labels = labels[idx[10000:]]
```

   Make sure you have version 1.0.2 or above of scikit-learn installed or the code may not work properly!

   It is recommended to use `numpy` and `scipy` where possible for speed, especially in distance computations. It is also highly recommended that you debug your code using a much smaller training set (e.g. 100 examples), and then run it on the larger training set once you're sure your code works properly.

   The k-NN algorithm is the first (and most trivial) classification algorithm we encounter in the course. In order to classify a new data point, it finds the k nearest neighbors of that point in the dataset and classifies according to the majority label. More details can be found on Wikipedia.

(a) Write a function that accepts as input: (i) a set of train images; (ii) a vector of labels, corresponding to the images; (iii) a query image; and (iv) a number k. The function will implement the k-NN algorithm to return a prediction of the query image, given the train images and labels. The function will use the k nearest neighbors, using the Euclidean L2 metric. In case of a tie between the k labels of neighbors, it will choose an arbitrary option.

(b) Run the algorithm using the first $n = 1000$ training images, on each of the test images, using $k = 10$. What is the accuracy of the prediction (i.e. the percentage of correct classifications)? What would you expect from a completely random predictor?

(c) Plot the prediction accuracy as a function of k, for $k = 1, \ldots, 100$ and $n = 1000$. Discuss the results. What is the best k?

(d) Using $k = 1$, run the algorithm on an increasing number of training images. Plot the prediction accuracy as a function of $n = 100, 200, \ldots, 5000$. Discuss the results.