CS480/680, Spring 2023

Assignment 1

Designer: Haochen Sun; Instructor: Hongyang Zhang

Released: May 15; Due: June 4, noon

Instructions

- We do not accept hand-written submissions.
- This assignment is due by noon on June 4, 2023.
- For questions labelled as "**coding**", please follow the instructions provided and implement the required features. The skeleton code is provided. Unless otherwise specified, all implementations should be in *Python* using a *Jupyter Notebook*. Before submission, please make sure that your code can run without any errors. Also, be sure to save the output of each cell, as any missing output may not be graded.
- Please submit the following TWO files to LEARN:
  - A write-up in PDF format: the written answers to ALL questions, including the reported results and plots of coding questions, in a single PDF file.
  - An IPYNB file: your implementations for ALL coding questions. Please save the output of each cell, or your coding questions may NOT be graded.

Question 1: Perceptron (30 points)

1. [**15 points**] Consider the AND, OR and XOR datasets, each of which labels all two-dimensional binary inputs $\left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix}\right\}$:

	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}0\\1\end{pmatrix}$	$\begin{pmatrix}1\\0\end{pmatrix}$	$\begin{pmatrix}1\\1\end{pmatrix}$
AND	0	0	0	1
OR	0	1	1	1
XOR	0	1	1	0

Table 1: The AND, OR and XOR datasets

For each dataset, prove or disprove if it is linearly separable. For a linearly separable dataset, write down the separating hyperplane. For a non-linearly separable dataset, argue that a separating hyperplane does not exist.

Solution:

For AND dataset, it is linearly separable, and one of the separating hyper planes is

$$\omega = \begin{pmatrix}1\\1\end{pmatrix}$$

and

$$b = -1.5$$

For OR dataset, it is linearly separable, and one of the separating hyper planes is

$$\omega = \begin{pmatrix}1\\1\end{pmatrix}$$

and

$$b = -0.5$$

For XOR dataset, it is not linearly separable, and the following is the proof:

Suppose the dataset is separable, thus there exist ω and b such that for all points, $sign(< x, \omega > +b)$ matches the data.

Note that

$$\begin{pmatrix}1\\0\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}1\\1\end{pmatrix}$$

so we have that there exists ω and b such that

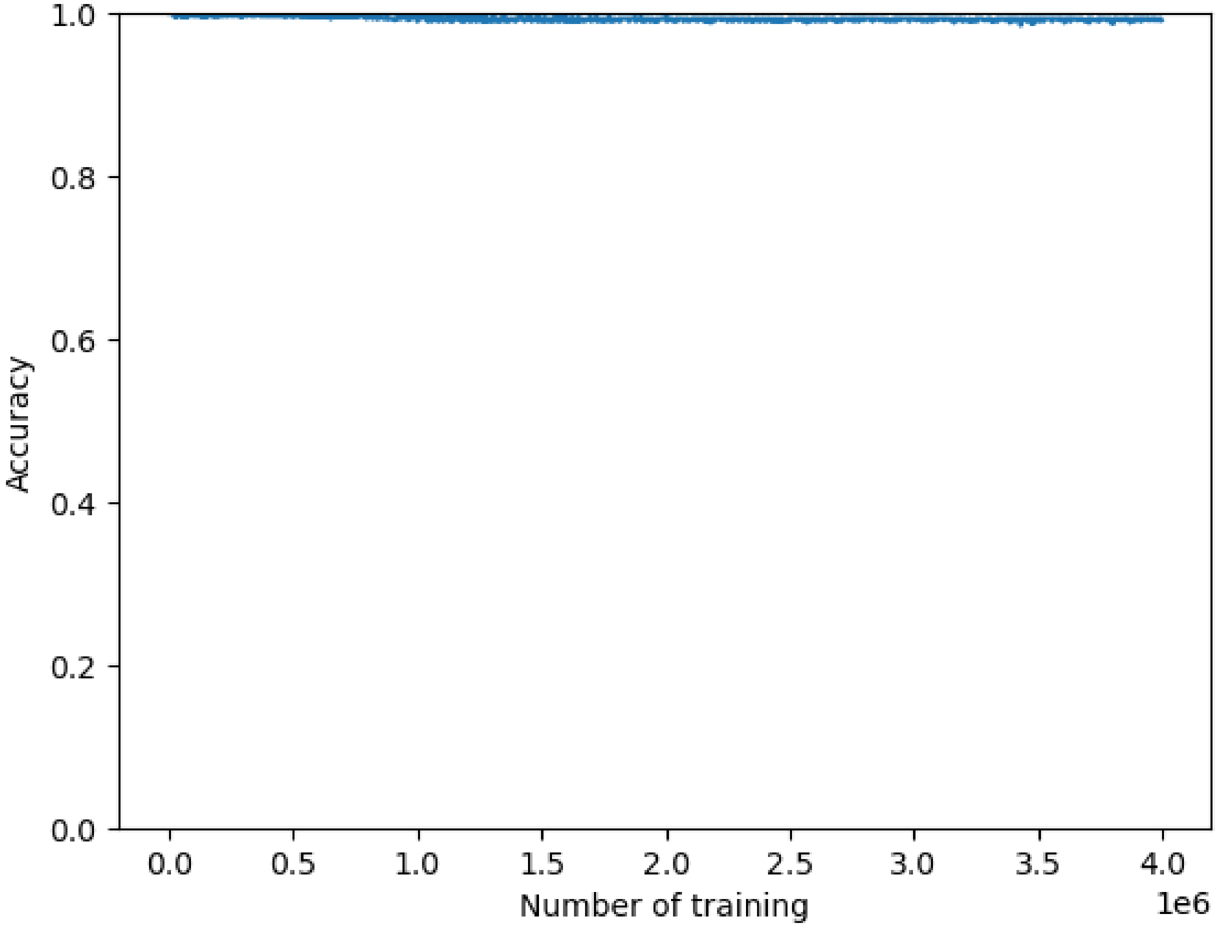
$$< \begin{pmatrix}1\\0\end{pmatrix}, \omega > +b + < \begin{pmatrix}0\\1\end{pmatrix}, \omega > +b =< \begin{pmatrix}1\\1\end{pmatrix}, \omega > +b$$

But, from data set, the sign of RHS should be different from the sign of LHS. Which is a contradiction.

Therefore, XOR dataset is not linearly separable.

2. [**15 points, coding**] Implement the perceptron algorithm on the Spambase dataset in *Python* using *Jupyter Notebook*. You may use the provided skeleton code, which downloads and pre-processes the dataset. Note that the target variable to be predicted is the last feature, `is_spam`. Plot the accuracy against the number of training steps and report the final accuracy.

It seems that the final accuracy is fluctuated around 1.


1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e6
Number of training

Question 2: Generalized Linear Models (40 points)

In class, we have discussed linear regression and logistic regression. While these models are useful for specific types of data, they belong to a broader class of models known as generalized linear models (GLMs). GLMs are models for data where the response variable y follows a distribution from the exponential family, and the mean of the response variable is related to the predictors via a link function. The GLM has the following form:

$$p(y|\mathbf{x}, \mathbf{w}, \tau) = h(y, \tau) \exp\left[\frac{y\eta - A(\eta)}{d(\tau)}\right], \tag{1}$$

where p is the probability mass function or probability density function of y, $\eta := \mathbf{w}^\top \mathbf{x}$ is the natural parameter, $A(\eta)$ is the log normalizer, and τ is the dispersion parameter, such that $d(\tau) > 0$ typically related to the variance of the conditional distribution. $h(y, \tau)$ is a normalization factor such that $p(y|\mathbf{x}, \mathbf{w}, \tau)$ sums/integrates to 1 over y. We denote the mapping from the linear input $\eta = \mathbf{w}^\top \mathbf{x}$ to the conditional expectation of y (i.e., $\mu = \mathbb{E}[y|\mathbf{x}, \mathbf{w}, \tau]$) as $\mu = \ell^{-1}(\eta)$, where ℓ is known as the link function, and ℓ^{-1} is known as the mean function. (For simplicity, omit the bias term throughout this question.)

1. **[10 points]** Show that linear regression and logistic regression are special cases of the GLM in (1). Identify $h(y, \tau)$, $A(\eta)$, $d(\tau)$, and the link function ℓ for each case.

   - Comparing the probability mass function of linear regression to the GLM, it is easy to notice that $h(y, \tau) = \frac{1}{\sqrt{2\pi\sigma^2}}$ and $\frac{y\eta - A(\eta)}{d(\tau)} = -\frac{1}{2\sigma^2}(y - \eta)^2$
   - Comparing the later equation, also notice that $d(\tau) > 0$, we have $d(\tau) = 2\sigma^2$, $A(\eta) = y^2 + \eta^2 - y\eta$. Notice that $\tau = \sigma^2$, we then have $h(y, \tau) = \frac{1}{\sqrt{2\pi\tau}}$, $d(\tau) = 2\tau$.
   - Rearrange the form of logistic regression, we have
   - $p(y|x, w) = \frac{(e^{-\eta})^{1-y}}{1+e^{-\eta}} = \frac{e^{-\eta}}{(1+e^{-\eta})e^{-y\eta}} = \frac{1}{1+e^{-\eta}} exp(y\eta - \eta)$
   - By comparing, we get $h(y, \tau) = \frac{1}{e^{-\eta}}$, $A(\eta) = \eta$, and $d(\tau) = 1$.

2. **[6 points]** Consider the Poisson regression model, which is defined by the following probability mass function:

   $$p(y|\mathbf{x}, \mathbf{w}) = \text{Pois}(y| \exp(\mathbf{w}^\top \mathbf{x})).$$

   Here, $\text{Pois}(y|\mu) = \frac{\mu^y e^{-\mu}}{y!}$ is the Poisson distribution parameterized by μ, with support $y \in \mathbb{N}$. Show that Poisson regression belongs to the family of GLMs defined in Equation (1), and identify $A(\eta)$ for this model.

   - Rearrange the expression, we have $p(y|x, w) = \frac{e^{y\eta} e^{-e^\eta}}{y!} = \frac{1}{y!} exp(y\eta - e^y)$
   - Then by comparing, we have $A(\eta) = e^y$

3. **[8 points]** Consider a dataset $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$, where $\mathbf{x}_n \in \mathbb{R}^d$ and y_n is the response variable. Derive the negative log-likelihood (NLL) for the GLM on this dataset based on Equation (1). Explain how the derived NLL is equivalent to the square loss (for linear regression) and binary cross entropy loss (for logistic regression) for finding the optimal $\mathbf{w}$.

   - $-log(P(y|x, w, \tau)) = -log(h(y, \tau) exp(\frac{y\eta - A(\eta)}{d(\tau)})) = -log(h(y, \tau)) - \frac{y\eta - A(\eta)}{d(\tau)}$

4. **[6 points]** Derive the negative log-likelihood (NLL) loss function for the Poisson regression model using the results from sub-questions 2 and 3. Simplify the expression as much as possible.

   - From 2, we have $A(\eta) = e^y$, $d(\tau) = 1$, $h(y, \tau) = \frac{1}{y!}$
   - Sub these into the form we have in 3, we get the NLL is $-log(\frac{1}{y!}) - y\eta + e^y$

5. **[10 points]** Prove that in Equation (1), $\mathbb{E}\left[y|\mathbf{x},\mathbf{w},\tau\right] = A'(\eta)$ and $\mathrm{Var}\left[y|\mathbf{x},\mathbf{w},\tau\right] = A''(\eta)d(\tau)$. (Hint: you may either assume that y is discrete or continuous. As the first step, sum/integrate both sides of Equation (1) over y. You may switch the order of the summation/integration and the derivative without justification.)

   - Assume it is continuous.
   - $\int h(y,\tau)exp[\frac{y\eta-A(\eta)}{d(\tau)}]dy$

Question 3: SVM kernels (30 points)

1. [**10 points**] Given a natural number M, consider $X = \{0, 1, \dots M\}$. Define $K(x, x') = \min\{x, x'\}$. Find a mapping $\phi : X \to \mathbb{R}^M$ such that for all $x, x' \in X$, $K(x, x') = \langle \phi(x), \phi(x') \rangle$.

   - Assume $\phi(x) = (\phi_1(x), \phi_2(x), ..., \phi_M(x))$
   - Then $< \phi(x), \phi(x') >= \sum_{i=1}^{M} \phi_i(xx')$

2. [**10 points**] Show that there exists a Hilbert space H and a mapping $\phi : \mathbb{R}^n \to H$ such that

   $$\langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle = 4 \langle \mathbf{u}, \mathbf{v} \rangle^2 + \langle \mathbf{u}, \mathbf{v} \rangle^3$$

   for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. (Hint: consider $H = \mathbb{R}^{n^2+n^3}$.)

   - Let $\phi(x) = ((x \bigotimes x \bigotimes x, x \bigotimes x))$
   - Then $< \phi(u), \phi(v) >=< (u \bigotimes u \bigotimes u, u \bigotimes u), (v \bigotimes v \bigotimes v, v \bigotimes v) >$
     $=< u \bigotimes u \bigotimes u, v \bigotimes v \bigotimes v > + < u \bigotimes u, v \bigotimes v >$
     $=< u \bigotimes u \bigotimes u, v \bigotimes v \bigotimes v > + \sum_i \sum_j u_i u_j v_i v_j =< u \bigotimes u \bigotimes u, v \bigotimes v \bigotimes v > + (\sum u_i v_i)^2$

3. [**10 points**] More generally, consider a polynomial f with non-negative coefficients, and construct H and ϕ such that

   $$\langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle = f(\langle \mathbf{u}, \mathbf{v} \rangle)$$

   for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$.

   - Let $f = \sum a_i x^i$
   - Define $\phi(x) = (\sqrt{a_0} x_0, \sqrt{a_1} x_1, ..., \sqrt{a_{n-1}} x_{n-1})$, note that $\phi : \mathbb{R}^n \to \mathbb{R}^n$
   - Then we have $< \phi(u), \phi(v) >=< (\sqrt{a_0} u_0, \sqrt{a_1} u_1, ..., \sqrt{a_{n-1}} u_{n-1}), (\sqrt{a_0} v_0, \sqrt{a_1} v_1, ..., \sqrt{a_{n-1}} v_{n-1}) >$

     $= a_0 u_0 v_0 + a_1 u_1 v_1 + ... + a_{n-1} u_{n-1} v_{n-1} = f(< u, v >)$