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. AND **is** linearly separable: $x_1 + x_2 - 1.5 = 0$;

OR **is** linearly separable: $x_1 + x_2 - 0.5 = 0$;

XOR **is not** linearly separable. Assume otherwise, such that a separating hyperplane $w_1x_1+w_2x_2+b = 0$ exists. Therefore, on the one hand, by the labelling of $\begin{pmatrix}0\\0\end{pmatrix}$ and $\begin{pmatrix}1\\1\end{pmatrix}$,

$$0 > b + w_1 + w_2 + b = w_1 + w_2 + 2b;$$

on the other hand, by the labelling of $\begin{pmatrix}0\\1\end{pmatrix}$ and $\begin{pmatrix}1\\0\end{pmatrix}$,

$$0 \leq w_2 + b + w_1 + b = w_1 + w_2 + 2b,$$

which is a contradiction. ■

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.

Solution. See the sample solution. Credit to Leo Feng. ■

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, which is 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.

   Proof. For linear regression,

$$\begin{aligned} p\left(y|\mathbf{x}, \mathbf{w}, \sigma^2\right) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - \mathbf{w}^\top \mathbf{x})^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{y^2}{2\sigma^2}\right) \exp\left(\frac{y\eta - \frac{1}{2}\eta^2}{\sigma^2}\right), \end{aligned}$$

   by setting $\tau = \sigma^2$, $h(y, \tau) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{y^2}{2\sigma^2}\right)$, $d(\sigma^2) = \sigma^2$ and $A(\eta) = \frac{1}{2}\eta^2$, linear regression satisfies the definition of (1). The link function $\ell(\mu)$ is the identity.

   For logistic regression, denote $\mu = \sigma(\eta) = \sigma(\mathbf{w}^\top \mathbf{x})$ where σ is the sigmoid function,

$$\begin{aligned} p\left(y|\mathbf{x}, \mathbf{w}\right) &= (1 - \mu)^{(1-y)} \mu^y \\ &= \exp\left[(1 - y)\log(1 - \mu) + y \log \mu\right] \\ &= \exp\left[y \log \frac{\mu}{1 - \mu} + \log(1 - \mu)\right] \\ &= \exp\left[y\eta - \log(1 + e^\eta)\right]. \end{aligned}$$

   Therefore, by setting $h(y, \tau) = 1, A(\eta) = \log(1 + e^\eta), d(\tau) = 1$, logistic regression satisfies the definition of (1). The link function $\ell(\mu) = \log \frac{\mu}{1-\mu}$. □

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

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

   Here, $\mathrm{Pois}(y|\mu) = \frac{\mu^y e^{-\mu}}{y!}$ 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.

Proof. For simplicity, denote $\mu = \exp(\eta) = \exp(\mathbf{w}^\top \mathbf{x})$. Therefore,

$$\begin{aligned} p\left(y|\mathbf{x}, \mathbf{w}\right) &= \frac{1}{y!} \exp\left(y \log \mu - \mu\right) \\ &= \frac{1}{y!} \exp\left(y\eta - e^\eta\right). \end{aligned}$$

Therefore, Poisson regression belongs to the GLM, and $A(\eta) = e^\eta$. □

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 $\mathbf{w}$.

   Solution. By definition, the NLL of (1) can be written as

$$-\sum_{n=1}^N p\left(y_n|\mathbf{x}_n, \mathbf{w}, \tau\right) = \frac{1}{d(\tau)} \sum_{n=1}^N \left(A(\mathbf{w}^\top \mathbf{x}_n) - y_n \mathbf{w}^\top \mathbf{x}_n\right) - Nh(y, \tau)$$

   For linear regression, minimizing the above expression over $\mathbf{w}$ is equivalent to minimizing

$$\sum_{n=1}^N \left(\frac{1}{2}(\mathbf{w}^\top \mathbf{x}_n)^2 - y_n \mathbf{w}^\top \mathbf{x}_n\right) = \sum_{n=1}^N \left(\frac{1}{2}(\mathbf{w}^\top \mathbf{x}_n)^2 - y_n \mathbf{w}^\top \mathbf{x}_n + \frac{1}{2} y_n^2\right) = \frac{1}{2} \sum_{n=1}^N \left(\mathbf{w}^\top \mathbf{x}_n - y_n\right)^2,$$

   which is the square loss. On the other hand, for logistic regression, denote $\hat{y}_n = \sigma(\mathbf{w}^\top \mathbf{x})$, then $\log \frac{\hat{y}_n}{1-\hat{y}_n} = \mathbf{w}^\top \mathbf{x}_n$. Therefore, minimizing the above express over $\mathbf{w}$ is equivalent to minimizing

$$\sum_{n=1}^N \left(-\log \frac{1}{1+e^{\mathbf{w}^\top \mathbf{x}_n}} - y_n \mathbf{w}^\top \mathbf{x}_n\right) = -\log(1-\hat{y}_n) - y_n \log \frac{\hat{y}_n}{1-\hat{y}_n} = -y_n \log \hat{y}_n - (1-y_n)\log(1-\hat{y}_n),$$

   which is the binary cross entropy loss defined. ■

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.

   Solution. Denote the predicted mean $\hat{y}_n := \exp(\mathbf{w}^\top \mathbf{x})$, then the loss function can be designed as $\sum_{n=1}^N (\hat{y}_n - y_n \log \hat{y}_n)$. ■

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.)

   Proof. We assume that y is a continuous random variable (e.g., linear regression), such that

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

   Taking derivatives on both sides w.r.t. η, we have

$$0 = \int_{y\in\mathbb{R}} h(y, \tau) \frac{y - A'(\eta)}{d(\tau)} \exp\left[\frac{y\eta - A(\eta)}{d(\tau)}\right] dy.$$

Therefore,

$$A'(\eta)\int_{y\in\mathbb{R}} h(y,\tau)\exp\left[\frac{y\eta - A(\eta)}{d(\tau)}\right]\,dy = \int_{y\in\mathbb{R}} h(y,\tau)y\exp\left[\frac{y\eta - A(\eta)}{d(\tau)}\right]\,dy = \mathbb{E}\left[y|\mathbf{x},\mathbf{w},\tau\right]\,dy.$$

On the other hand, the LHS of the equation above equals $A'(\eta)$ since $\int_{y\in\mathbb{R}} h(y,\tau)\exp\left[\frac{y\eta-A(\eta)}{d(\tau)}\right]\,dy = 1$, thus completing the proof of the first statement. Additionally, by taking the derivative of η in the equation above, we have

$$A''(\eta) = \int_{y\in\mathbb{R}} h(y,\tau)y\frac{y - A'(\eta)}{d(\tau)}\exp\left[\frac{y\eta - A(\eta)}{d(\tau)}\right]\,dy,$$

such that

$$d(\tau)A''(\eta) = \mathbb{E}\left[y^2|\mathbf{x},\mathbf{w},\tau\right] - A'(\eta)\mathbb{E}\left[y|\mathbf{x},\mathbf{w},\tau\right] = \mathbb{E}\left[y^2|\mathbf{x},\mathbf{w},\tau\right] - \mathbb{E}\left[y|\mathbf{x},\mathbf{w},\tau\right]^2 = \mathrm{Var}\left[y|\mathbf{x},\mathbf{w},\tau\right],$$

which completes the proof of the second statement. □

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 $x, x' \in X$, $K(x, x') = \langle \phi(x), \phi(x') \rangle$.

   Solution. $\phi(x) = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$, where the first x dimensions are 1 and the other $M - x$ dimensions are 0. ■

2. **[10 points]** Show that there exist 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 $H = \mathbb{R}^{n^2+n^3}$.)

   Solution. See subquestion 3. ■

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

   Proof. Write $f(z) = c_0 + c_1 z + \cdots + c_K z^K$, where $c_0, c_1, \dots c_K \geq 0$, and $\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$, $\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$ Then for each $1 \leq k \leq K$, note that

   $$\begin{aligned} \langle \mathbf{u}, \mathbf{v} \rangle^k &= \left(\sum_{i=1}^{n} u_i v_i \right)^k \\ &= \sum_{1 \leq i_1, i_2, \dots i_k \leq n} u_{i_1} v_{i_1} u_{i_2} v_{i_2} \dots u_{i_k} v_{i_k} \\ &= \sum_{1 \leq i_1, i_2, \dots i_k \leq n} (u_{i_1} u_{i_2} \dots u_{i_k})(v_{i_1} v_{i_2} \dots v_{i_k}). \end{aligned}$$

   Therefore, by defining $\phi_k : \mathbb{R}^n \to \mathbb{R}^{n^k}$ such that $\phi_k(\mathbf{u}) = (u_{i_1} u_{i_2} \dots u_{i_k})_{1 \leq i_1, i_2, \dots i_k \leq n}$, $\langle \mathbf{u}, \mathbf{v} \rangle^k = \langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle$. Therefore, by defining

   $$\phi(\mathbf{u}) = \begin{pmatrix} \sqrt{c_0} \\ \sqrt{c_1}\phi_1(\mathbf{u}) \\ \sqrt{c_2}\phi_2(\mathbf{u}) \\ \vdots \\ \sqrt{c_K}\phi_K(\mathbf{u}) \end{pmatrix},$$

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