CS 446 / ECE 449 — Homework 2

yiminc2

Version 1.0

Instructions.

- Homework is due **Wednesday, October 1^{st}, 11:59 a.m**; you have **3** late days in total for **all Homeworks**.
- Everyone must submit **individually** at gradescope under `hw2` and `hw2code`.
- The "written" submission at `hw2` **must be typed**, and submitted in any format gradescope accepts (to be safe, submit a PDF). You may use LaTeX, markdown, google docs, MS word, whatever you like; but it must be typed!
- When submitting at `hw2`, gradescope will ask you to **mark out boxes around each of your answers**; please do this precisely!
- Please make sure your NetID is clear and large on the first page of the homework.
- Your solution **must** be written in your own words. Please see the course webpage for full **academic integrity** information. You should cite any external reference you use.
- We reserve the right to reduce the auto-graded score for `hw2code` if we detect funny business (e.g., your solution lacks any algorithm and hard-code answers you obtained from someone else, or simply via trial-and-error with the autograder).
- When submitting to `HW2 coding`, only upload `hw2_q2.py`, `hw2_q4.py` and `hw2_utils.py`. Additional files will be ignored. (**DO NOT change the name of these three files!**)

1. Naive Bayes (25 pt)

(a) In Naive Bayes classification, the number of parameters to be estimated for a Bayesian classifier is reduced by assuming conditional independence when modeling $P(X|Y)$. Conditional independence is defined as follows:

Definition: Let X, Y, and Z be random variables. We say that X is conditionally independent of Y given Z, written as $X \perp Y|Z$, if and only if:

$$P(X = x_i|Y = y_j, Z = z_k) = P(X = x_i|Z = z_k), \forall i, j, k$$

Given this definition, please answer the following questions:

i. If $X \perp Y|Z$, can we conclude that $P(X, Y|Z) = P(X|Z)P(Y|Z)$? Explain your reasoning. (2 pt)

ii. If $X \perp Y|Z$, can we conclude that $P(X, Y) = P(X)P(Y)$? Explain your reasoning. (2 pt)

iii. Suppose X is a vector of d Boolean attributes, and Y is a discrete variable that takes C possible values. Let $\theta_{jc} = P(X_j|Y = c)$. How many independent θ_{jc} parameters must be estimated? (2 pt)

iv. Now suppose X is a vector of d real-valued attributes, and each X_j follows a Normal (Gaussian) distribution: $P(X_j = x_j|Y = c) = \mathcal{N}(x_j|\mu_{jc}, \sigma_{jc})$. How many distinct μ_{jc} and σ_{jc} parameters must be estimated? (2 pt)

v. We can write the classification rule for Naive Bayes as:

$$Y^{new} \leftarrow \arg\max_c \frac{P(Y = c)\prod_{j=1}^{d} P(X_j^{new}|Y = c)}{\sum_{v=1}^{C} P(Y = v)\prod_{k=1}^{d} P(X_k^{new}|Y = v)}$$

When estimating Y, we often omit the denominator in the calculation. Why is it acceptable to do this? (2 pt)

vi. Is it possible to compute $P(X)$ using the parameters estimated in Naive Bayes? (2 pt)

(b) Consider a classification problem where the input vector $X =< X_1, X_2, X_3 >$ consists of three boolean features $X_j \in \{0, 1\}, j = \{1, 2, 3\}$ and the label $Y \in \{0, 1\}$. You are given a dataset of N i.i.d. labeled examples $\{X^{(i)}, y^{(i)}\}_{i=1}^{N}$. For X_1, X_2 and X_3 we have: $P(X_1|X_2, X_3, Y) = P(X_1|Y)$, $P(X_2|X_1, X_3, Y) = P(X_2|Y)$ and $P(X_3|X_1, X_2, Y) = P(X_3|X_1)$.

i. Express the joint distribution $P(X_1, X_2, X_3, Y)$ as a product of simpler conditional probabilities, i.e. each variable depends only on atmost one another variable. (2 pt)

ii. Derive the maximum likelihood estimators for the following quantities (6 pt, 2pt each) :

A. $P(Y = 1)$

B. $P(X_1 = 1 \mid Y = y)$, for $y \in \{0, 1\}$.

C. $P(X_3 = 1 \mid Y = y)$, for $y \in \{0, 1\}$.

Note: It is sufficient to leave your answers as sums of indicator functions or fractions.

(c) Consider the same conditional independence structure as in Part (b), but this time with continuous features X_j drawn from Gaussian distributions instead of boolean. Specifically, let $P(Y = 0) = P(Y = 1) = 0.5$ and

$$\begin{aligned}
&(X_1 \mid Y = y) \sim \mathcal{N}(\mu_{1y}, 1), && \mu_{10} = 0,\ \mu_{11} = 1,\\
&(X_2 \mid Y = y) \sim \mathcal{N}(\mu_{2y}, 1), && \mu_{20} = 0,\ \mu_{21} = 1,\\
&(X_3 \mid X_1 = x_1) \sim \mathcal{N}(2x_1, 1) && \text{(independent of } Y \text{ given } X_1).
\end{aligned}$$

i. Derive the MAP rule for a new point X^{new}. (2 pt)

ii. Using your classification rule, classify the following two points:

$$X^{(a)} =< 0.2, 0.7, -10 >, \qquad X^{(b)} =< 0.2, 0.7, 10 > .$$

Do the predicted labels differ between these two cases? Explain why or why not. (3 pt)

(a) i. If $X \perp Y \mid Z$, can we conclude that $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$?
Ans: Yes

$$\begin{aligned} P(X, Y \mid Z) &= \frac{P(X, Y, Z)}{P(Z)} \\ &= \frac{P(Z)P(X, Y \mid Z)}{P(Z)} \\ &= \frac{P(Z)P(Y \mid Z)P(X \mid Y, Z)}{P(Z)} \quad \text{(chain rule)} \\ &= \frac{P(Z)}{P}(Y \mid Z)P(X \mid Z)P(Z) \quad (Y \perp X \mid Z) \\ &= P(Y \mid Z)P(X \mid Z) \end{aligned}$$

ii. If $X \perp Y \mid Z$, can we conclude that $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$?
Ans: No

$\because P(X, Y) = P(X)P(Y) \iff X \perp Y$, and we only know that $X \perp Y \mid Z$
$\therefore P(X, Y) \neq P(X)P(Y)$

iii. How many independent θ_{jc} parameters must be estimated?
Ans: C $\times$ d indepedent θ_{jc} parameters

- X has d attributes.
- For each class c, we estimate θ_{jc} for each attribute X_j.
- Therefore, the total number of independent parameters is $C \times d$.

iv. How many distinct μ_{jc} and σ_{jc} parameters must be estimated?
Ans: $2 \times C \times d$ distinct μ_{jc} and σ_{jc} parameters with both having the same number $C \times d$

- Each attribute X_j requires a distinct set of μ_{jc} and σ_{jc}
- The respective numbers of μ_{jc} and σ_{jc} are both $C \times d$
- Therefore, the total number of distinct parameters are $2 \times C \times d$

v. Why is omitting the denominator in the calculation acceptable for the estimation of Y.
Ans: The Y variable is not dependent on the denominator as the denominator is just a constant for normalization.

vi. Is it possible to calculate the $P(X)$ with Naive Bayes?
Ans: Yes.

- Given $P(X = x) = \sum_{c=1}^{C} P(X = x \mid Y = c)P(Y = c)$
- and the likelihood of Naive Bayes: $P(X \mid Y) = \prod_{j=1}^{d} P(X_j \mid Y = c)$,
- we can derive that $P(X) = \sum_{c=1}^{C} P(Y = c) \prod_{j=1}^{d} P(X_j \mid Y = c)$

(b) i. Express the joint distribution $P(X_1, X_2, X_3, Y)$ as a product of simpler conditional probabilty.
Ans: $P(Y)P(X_1 \mid Y)P(X_3)P(X_2 \mid Y)$

- $P(X_1 \mid X_2, X_3, Y) = P(X_1 \mid Y) \Rightarrow X_1 \perp X_2, X_3 \mid Y$
- $P(X_2 \mid X_1, X_3, Y) = P(X_2 \mid Y) \Rightarrow X_2 \perp X_1, X_3 \mid Y$
- $P(X_3 \mid X_1, X_2, Y) = P(X_3 \mid X_1) \Rightarrow X_3 \perp X_2, Y \mid X_1$

$$\begin{aligned} P(X_1, X_2, X_3, Y) &= P(Y)P(X_1, X_2, X_3 \mid Y) \\ &= P(Y)P(X_1 \mid Y)P(X_2, X_3 \mid X_1, Y) \\ &= P(Y)P(X_1 \mid Y)\underbrace{P(X_3 \mid X_1, Y)}_{=P(X_3|X_1)}\underbrace{P(X_2 \mid X_3, X_1, Y)}_{=P(X_2|Y)} \\ &= P(Y)P(X_1 \mid Y)P(X_3 \mid X_1)P(X_2 \mid Y) \end{aligned}$$

ii. Derive the maximum likelihood estimators for the following quantities:

Assume that $\mathbb{I}$ is the indicator function s.t. $\mathbb{I}\{e\} = \begin{cases} 1 & \text{(e is true)} \\ 0 & \text{(e is wrong)} \end{cases}$

A. P(Y = 1)
Ans: $\frac{\sum_{i=1}^{N} \mathbb{I}\{y^{(i)}=1\}}{N}$

B. $P(X_1 = 1 \mid Y = y)$, for $y \in \{0, 1\}$
Ans: $\frac{\sum_{i=1}^{N} \mathbb{I}\{x_1^{(i)}=1, y^{(i)}=y\}}{\sum_{i=1}^{N} \mathbb{I}\{y^{(i)}=y\}}$

C. $P(X_3 = 1 \mid Y = y)$, for $y \in \{0, 1\}$
Ans: $\sum_{x_1=0}^{1} \frac{\sum_{i=0}^{N} \mathbb{I}\{x_1^{(i)}=x_1, x_3^{(i)}=1\}}{\sum_{i=0}^{N} \mathbb{I}\{x_1^{(i)}=x_1\}} \frac{\sum_{i=1}^{N} \mathbb{I}\{y^{(i)}=y, x_1^{(i)}=x_1\}}{\sum_{i=1}^{N} \mathbb{I}\{y^{(i)}=y\}}$

$$P(X_3 \mid Y) = \sum_{x_1=0}^{1} P(X_3, X_1 = x_1 \mid Y) \quad (Marginalization)$$

$$P(X_3 \mid Y) = \sum_{x_1=0}^{1} P(X_3 \mid X_1 = x_1, Y)P(X_1 = x_1 \mid Y)$$

$$\Rightarrow P(X_3 = 1 \mid Y = y) = \sum_{x_1=0}^{1} P(X_3 = 1 \mid X_1 = x_1)P(X_1 = x_1 \mid Y)$$

$$\Rightarrow P(X_3 = 1 \mid Y = y) = \sum_{x_1=0}^{1} \frac{\sum_{i=0}^{N} \mathbb{I}\{x_1^{(i)} = x_1, x_3^{(i)} = 1\}}{\sum_{i=0}^{N} \mathbb{I}\{x_1^{(i)} = x_1\}} \frac{\sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = y, x_1^{(i)} = x_1\}}{\sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = y\}}$$

(c) i. Derive the MAP for a new point X^{new}

Ans: $\begin{cases} 1 & (x_1 + x_2 > 1) \\ 0 & (otherwise) \end{cases}$

Given the same conditional independence structure

$$\begin{aligned} &P(X_1, X_2, X_3, Y) = P(Y)P(X_1 \mid Y)P(X_2 \mid Y)P(X_3 \mid X_1) \\ &\Rightarrow P(X_1, X_2, X_3 \mid Y)P(Y) = P(Y)P(X_1 \mid Y)P(X_2 \mid Y)P(X_3 \mid X_1) \\ &\Rightarrow P(X_1, X_2, X_3 \mid Y) = P(X_1 \mid Y)P(X_2 \mid Y)P(X_3 \mid X_1) \end{aligned}$$

Under the condition where Y = 1 is chosen:

$$
\begin{aligned}
&P(Y=1 \mid X_1, X_2, X_3) > P(Y=0 \mid X_1, X_2, X_3) \\
&\iff P(Y=1)\, P(X_1, X_2, X_3 \mid Y=1) > P(Y=0)\, P(X_1, X_2, X_3 \mid Y=0) \\
&\iff P(X_1 \mid Y=1) P(X_2 \mid Y=1) > P(X_1 \mid Y=0) P(X_2 \mid Y=0) \\
&\iff \log \frac{P(X_1 \mid Y=1)}{P(X_1 \mid Y=0)} + \log \frac{P(X_2 \mid Y=1)}{P(X_2 \mid Y=0)} > 0 \\
&\iff -\tfrac{1}{2}\Big[(x_1-\mu_{11})^2 - (x_1-\mu_{10})^2\Big] - \tfrac{1}{2}\Big[(x_2-\mu_{21})^2 - (x_2-\mu_{20})^2\Big] > 0 \\
&\iff (\mu_{11}-\mu_{10})x_1 - \tfrac{1}{2}(\mu_{11}^2 - \mu_{10}^2) + (\mu_{21}-\mu_{20})x_2 - \tfrac{1}{2}(\mu_{21}^2 - \mu_{20}^2) > 0
\end{aligned}
$$

Plug the values of μs in

$$x_1 - \tfrac{1}{2} + x_2 - \tfrac{1}{2} > 0 \qquad\qquad x_1 + x_2 > 1$$

Therefore, the MAP for a new point X^{new} is

$$\begin{cases} 1 & (x_1 + x_2 > 1) \\ 0 & (otherwise) \end{cases}$$

ii. Using your classification rule, classify the following two points:

$$X^{(a)} =< 0.2, 0.7, -10 >, \qquad X^{(b)} =< 0.2, 0.7, 10 >$$

Do the predicted labels differ between these two cases? Explain why?
Ans: Both $X^{(a)}$ and $X^{(b)}$ are classified with the label $y = 0$ as the sum of their X_1 and X_2 are equal (0.9) and their X_3 values are not used by the classifier.

2. Gaussian Naive Bayes. (25 pt)

Recall from Lecture 7 (slide 5), taking $\log(\cdot)$ of the objective function, we have the decision rule as :

$$Y^{\text{new}} = \arg\max_y \left(\left(\sum_{j=1}^{d} \log P(X_j^{\text{new}} | Y = y) \right) + \log P(Y = y) \right)$$

In a Gaussian Naive Bayes, the features X_j^{new} are continuous variables, and the probability $P(X_j^{\text{new}} | Y = y)$ is modeled as a Gaussian distribution

$$P(X_j^{\text{new}} | Y = y) = \mathcal{N}(X_j^{\text{new}} \mid \mu_{y,j}, \sigma_{y,j}^2)$$

For simplicity, we assume that there exists at least example that belong to each class.

- There are d attributes to describe each example.
- We use pair $(\boldsymbol{x}^{(i)}, y^{(i)})$ to represent ith example, where $\boldsymbol{x}^{(i)}$ is a length-d vector that describes its properties, and $y^{(i)}$ is a scalar representing its label.
- For $j \in \{1, 2, \cdots, d\}$, $x_j^{(i)} \in \mathbb{R}$ is the value of attribute j of each example i; $y^{(i)} = 0$ means example i is in class 0, and $y^{(i)} = 1$ means example i is in class 1.
- We use $(\boldsymbol{X}, \boldsymbol{y})$ to represent a dataset of size N, where $\boldsymbol{X} = \left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots, \boldsymbol{x}^{(N)} \right)^\top$ is a $N \times d$ matrix, and $\boldsymbol{y} = \left(y^{(1)}, y^{(2)}, \cdots, y^{(N)} \right)$ is a length-N vector.

You are given two datasets in this homework: $(\boldsymbol{X}_{\text{train}}, \boldsymbol{y}_{\text{train}})$ and $(\boldsymbol{X}_{\text{test}}, \boldsymbol{y}_{\text{test}})$, which can be obtained by calling `gaussian_dataset("train")` and `gaussian_dataset("test")` in `hw2_utils.py`. Your tasks are:

(a) Implement the function `gaussian_theta(X, y)` in `hw2_q2.py`.
The input is the training dataset $(\boldsymbol{X}_{\text{train}}, \boldsymbol{y}_{\text{train}})$. The output is $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. Both of them are $2 \times d$ matrices in PyTorch float tensor, where $\mu_{y,j}$ and $\sigma_{y,j}^2$ are Gaussian distribution parameters of the MLE estimation of $P(X_j = 1 \mid Y = y)$. (8 pt)

(b) Implement the function `gaussian_p(y)` in `hw2_q2.py`.
The input is the label part $\boldsymbol{y}_{\text{train}}$ of the training dataset. The output p is a scalar, which is the MLE for $P(Y = 0)$. (4 pt)

(c) Implement the function `gaussian_classify(mu, sigma2, p, X)` in `hw2_q2.py`.
The input is the output $\boldsymbol{\mu}, \boldsymbol{\sigma^2}, p$ from the two functions above, and the label part $\boldsymbol{X}_{\text{test}}$ of the testing dataset of size N. The output $\widehat{\boldsymbol{y}}$ is an length-N vector, where $\widehat{\boldsymbol{y}}^{(i)}$ is the predicted label (0 or 1) for the object with properties $\boldsymbol{x}_{\text{test}}^{(i)}$, or the i-th row of $\boldsymbol{X}_{\text{test}}$. For simplicity, it's guaranteed that there will be no ties (so the arg max will be unique). (6pt)
Note: Please use the logarithmic form of both $\widehat{Y}$ and Gaussian PDF to avoid precision issues.
Library routines: `torch.log`, `torch.sum`, `torch.mean`, `torch.var` (Please use `unbiased=False`).

(d) Show that, in Gaussian Naive Bayes with equal class priors $P(Y = 0) = P(Y = 1) = 0.5$ and with class-conditional distributions

$$(X_j \mid Y = y) \sim \mathcal{N}(\mu_{y,j}, \sigma_j^2),$$

where the variances σ_j^2 are the same across classes (but may differ across features), the MAP classification rule is equivalent to a *linear classifier* in $\boldsymbol{x} =< x_1, \ldots, x_d >$. That is, derive the explicit form of the decision boundary and show it can be written as

$$\widehat{y}(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \sum_{j=1}^{d} w_j x_j > \tau, \\ 0, & \text{if } \sum_{j=1}^{d} w_j x_j < \tau. \end{cases}$$

for some weights w_j and threshold τ depending on $(\mu_{0,j}, \mu_{1,j}, \sigma_j^2)$. (7pt)

(d) To label input data with 1, $P(Y = 1 \mid X_j) > P(Y = 0 \mid X_j)$ should hold.

$$
\begin{aligned}
&P(Y = 1 \mid x) > P(Y = 0 \mid x) \\
&\iff P(x \mid Y = 1) > P(x \mid Y = 0) \qquad \text{(equal priors)} \\
&\iff \prod_{j=1}^{d} P(X_j = x_j \mid Y = 1) > \prod_{j=1}^{d} P(X_j = x_j \mid Y = 0) \\
&\iff \sum_{j=1}^{d} \log P(X_j = x_j \mid Y = 1) > \sum_{j=1}^{d} \log P(X_j = x_j \mid Y = 0) \\
&\iff \sum_{j=1}^{d} \left[-\tfrac{1}{2} \log(2\pi\sigma_j^2) - \frac{(x_j - \mu_{1j})^2}{2\sigma_j^2} \right] > \sum_{j=1}^{d} \left[-\tfrac{1}{2} \log(2\pi\sigma_j^2) - \frac{(x_j - \mu_{0j})^2}{2\sigma_j^2} \right] \\
&\iff \sum_{j=1}^{d} \frac{(x_j - \mu_{0j})^2 - (x_j - \mu_{1j})^2}{2\sigma_j^2} > 0 \\
&\iff \sum_{j=1}^{d} \frac{2x_j(\mu_{1j} - \mu_{0j}) + (\mu_{0j}^2 - \mu_{1j}^2)}{2\sigma_j^2} > 0 \\
&\iff \sum_{j=1}^{d} \left(\frac{\mu_{1j} - \mu_{0j}}{\sigma_j^2} \right) x_j > -\frac{\mu_{0j}^2 - \mu_{1j}^2}{2\sigma_j^2} \\
&\text{Let } w_j = \left(\frac{\mu_{1j} - \mu_{0j}}{\sigma_j^2} \right), \tau = -\frac{\mu_{0j}^2 - \mu_{1j}^2}{2\sigma_j^2} \\
&\iff \sum_{j=1}^{d} w_j x_j > \tau
\end{aligned}
$$

$\because P(Y = 1 \mid x) > P(Y = 0 \mid x) \iff \sum_{j=1}^{d} w_j x_j > \tau$, and following the same proceudre,

we can derive $P(Y = 0 \mid x) > P(Y = 1 \mid x) \iff \sum_{j=1}^{d} w_j x_j < \tau$

$$
\therefore \hat{y}(x) = \begin{cases} 1 & (\sum_{j=1}^{d} w_j x_j > \tau) \\ 0 & (\sum_{j=1}^{d} w_j x_j < \tau) \end{cases}
$$

3. Logistic Regression. (25 pt)

In logistic regression, the class probability is modeled as

$$P(y^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}) = \sigma(y^{(i)}\mathbf{w}^\top \mathbf{x}^{(i)})$$

where $\mathbf{x}^{(i)} \in \mathbb{R}^{d+1}$ represents the feature vector, $y^{(i)} \in \{-1, +1\}$ is the class label, and $\sigma(s) = \frac{1}{1+e^{-s}}$ is the sigmoid function.

(Note: We exclude the bias term w_0 as it can be absorbed into the weight vector: $\mathbf{w} = \begin{bmatrix} \uparrow \\ \boldsymbol{w} \\ \downarrow \\ w_0 \end{bmatrix}$, and transform $\boldsymbol{x}$ to $\mathbf{x} = \begin{bmatrix} \uparrow \\ \boldsymbol{x} \\ \downarrow \\ 1 \end{bmatrix}$, which we discussed in Lecture 3 "notation hack".)

(a) Prove that the sigmoid function $\sigma(\cdot)$ satisfies the property

$$\sigma(-s) = 1 - \sigma(s)$$

By demonstrating this, show that, $P(y^{(i)} = -1|\boldsymbol{x}^{(i)}) + P(y^{(i)} = 1|\boldsymbol{x}^{(i)}) = 1$. (2 pt)

(b) Prove that

$$\sigma'(s) = \sigma(s)(1 - \sigma(s))$$

(2 pt)

(c) Derive the gradient of the log-likelihood function, that is, compute

$$\nabla_w \log P(\boldsymbol{y}|\mathbf{X}, \mathbf{w})$$

where $\boldsymbol{y} = [y^{(1)}, y^{(2)}, ..., y^{(N)}]^T$ and $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(N)}\}$. (5 pt)

(d) Derive the Hessian matrix $\boldsymbol{H}$ of the log-likelihood function, where each entry H_{ab} is given by

$$H_{ab} = \frac{\partial^2}{\partial w_a \partial w_b} \log P(\boldsymbol{y}|\mathbf{X}, \mathbf{w})$$

(5 pt)

(e) Prove that the Hessian is negative semi-definite, i.e., show that $\boldsymbol{z}^T \boldsymbol{H} \boldsymbol{z} \leq 0$ for any vector $\boldsymbol{z} \in \mathbb{R}^{d+1}$. By doing so, we can conclude that the log-likelihood is concave and has no local maxima, only a global maximum. (5 pt)

(f) **Gradient Descent Update Rule:** Use the gradient of the log-likelihood to derive the weight update rule for one iteration of gradient descent. Assume a learning rate α. (3 pt)

(g) **Newton's Method Update Rule:** Use Newton's method incorporating the Hessian matrix to derive the weight update rule. (3 pt)

(a) Prove that the sigmoid function satisfies the property $\sigma(-s) = 1 - \sigma(s)$, and show that $P(y^{(i)} = -1 \mid x^{(i)}) + P(y^{(i)} = 1 \mid x^{(i)}) = 1$.

$$
\begin{aligned}
\sigma(-s) &= \frac{1}{1+e^s} \\
&= \frac{1+e^s}{1+e^s} - \frac{e^s}{1+e^s} \\
&= 1 - \frac{e^s}{1+e^s} \\
&= 1 - \frac{1}{\frac{1}{e^s}+1} \\
&= 1 - \frac{1}{e^{-s}+1} \\
&= 1 - \sigma(s)
\end{aligned}
$$

Given this proof we can derive that:

$$
\begin{aligned}
&P(y^{(i)} = -1 \mid x^{(i)}) + P(y^{(i)} = 1 \mid x^{(i)}) \\
&= \sigma(-w^\intercal x^{(i)}) + \sigma(w^\intercal x^{(i)}) \\
&\because \sigma(-s) = 1 - \sigma(s) \\
&\therefore \sigma(-w^\intercal x^{(i)}) + \sigma(w^\intercal x^{(i)}) = 1 - \sigma(w^\intercal x) + \sigma(w^\intercal x) = 1
\end{aligned}
$$

(b) Prove that $\sigma'(s) = \sigma(s)(1 - \sigma(s))$

$$
\begin{aligned}
\sigma'(s) &= -(1+e^{-s})^{-2}(-e^{-s}) \\
&= (1+e^{-s})^{-1}(1+e^{-s})^{-1}(e^{-s}) \\
&= \sigma(s)\frac{e^{-s}}{1+e^{-s}} \\
&= \sigma(s) \underbrace{\frac{1}{e^s+1}}_{\sigma(-s)=1-\sigma(s)} \\
&= \sigma(s)(1-\sigma(s))
\end{aligned}
$$

(c) Derive the gradient of the log-likelihood function.
Ans: $X(\sigma(-y \odot Xw) \odot y)$

$$
\begin{aligned}
\log P(y \mid X, w) &= \sum_{i=1}^{N} \log P(y^{(i)} \mid x^{(i)}, w) \\
&= \sum_{i=1}^{N} \log \sigma(y^{(i)} w^\intercal x^{(i)}) \\
\nabla_w \log P(y \mid X, w) &= \sum_{i=1}^{N} (1 - \sigma(y^{(i)} w^\intercal x^{(i)})) y^{(i)} x^{(i)} \\
&= \sum_{i=1}^{N} \sigma(-y^{(i)} w^\intercal x^{(i)}) y^{(i)} x^{(i)}
\end{aligned}
$$

Given $y = \left[y^{(1)}, y^{(2)}, \ldots, y^{(N)}\right]^\intercal$ and $X = \left[x^{(1)}, x^{(2)}, \ldots, x^{(N)}\right]$, we rewrite the gradient in the matrix form:

$$\sum_{i=1}^{N} \sigma(-y^{(i)} w^\intercal x^{(i)}) y^{(i)} x^{(i)}$$
$$= X(\sigma(-y \odot Xw) \odot y) \qquad (Xc = \sum_{i=1}^{N} c_i x^{(i)} \text{ where } c = [c_1, c_2, \ldots, c_N]^\intercal)$$

(d) Derive the Hessian matrix H of the log-likelihood function.
Ans: $-XDX^\intercal$
Partially differentiate the previous result by w again

$$\frac{\partial}{\partial w} \sum_{i=1}^{N} \sigma(-y^{(i)} w^\intercal x^{(i)}) y^{(i)} x^{(i)}$$
$$= \sum_{i=1}^{N} \sigma(-y^{(i)} w^\intercal x^{(i)})(1 - \sigma(-y^{(i)} w^\intercal x^{(i)}))(-y^{(i)} x^{(i)})(y^{(i)} x^{(i)})$$
$$= \sum_{i=1}^{N} (1 - \sigma(y^{(i)} w^\intercal x^{(i)}))\sigma(y^{(i)} w^\intercal x^{(i)})(-y^{(i)} x^{(i)})(y^{(i)} x^{(i)})$$
$$= \sum_{i=1}^{N} (1 - \sigma(y^{(i)} w^\intercal x^{(i)}))\sigma(y^{(i)} w^\intercal x^{(i)})(-x^{(i)}) x^{(i)\intercal} \quad (y \in \{1, -1\})$$

Rewrite it into the matrix form

$$\sum_{i=1}^{N} (1 - \sigma(y^{(i)} w^\intercal x^{(i)}))\sigma(y^{(i)} w^\intercal x^{(i)})(-x^{(i)}) x^{(i)\intercal}$$
$$= -X \sum_{i=1}^{N} (1 - \sigma(y^{(i)} w^\intercal x^{(i)}))\sigma(y^{(i)} w^\intercal x^{(i)}) X^\intercal$$
$$\because (1 - \sigma(y^{(i)} w^\intercal x^{(i)}))\sigma(y^{(i)} w^\intercal x^{(i)}) \text{ produces scalar values}$$
$$\therefore \text{we can rewrite} \sum_{i=1}^{N} (1 - \sigma(y^{(i)} w^\intercal x^{(i)}))\sigma(y^{(i)} w^\intercal x^{(i)}) \text{as a diagonal matrix } D$$
$$= -XDX^\intercal$$

(e) Prove that the Hessian is negative semi-definite.

$$
\begin{aligned}
z^\intercal Hz &= -z^\intercal XDX^\intercal z \\
&= -(X^\intercal z)^\intercal D(X^\intercal z) \\
&\text{Let } a = X^\intercal z, \text{ and rewrite the expression in quardratic form} \\
&= -\sum_{i=1}^{N}(1-\sigma(y^{(i)}w^\intercal x^{(i)}))\sigma(y^{(i)}w^\intercal x^{(i)})a_i^2 \\
&\because (1-\sigma(y^{(i)}w^\intercal x^{(i)}))\sigma(y^{(i)}w^\intercal x^{(i)}) \text{ and } a_i^2 \text{ should always be greater than or equal to 0} \\
&\therefore -\sum_{i=1}^{N}(1-\sigma(y^{(i)}w^\intercal x^{(i)}))\sigma(y^{(i)}w^\intercal x^{(i)})a_i^2 \leq 0 \\
&\therefore z^\intercal Hz \text{ is negative semi-definite}
\end{aligned}
$$

(f) Use the gradient of the log-likelihood to derive the weight update rule for one iteration of gradient descent.
Ans: $w_{t+1} = w_t + \alpha X(\sigma(-y \odot Xw) \odot y)$ **at iteration** t

(g) Use Newton's method incoporating the Hessian matrix to derive the weight update rule. **Ans:** $w_{t+1} = w_t + (XDX^\intercal)^{-1}\nabla l(w_t)$

To maximize the log-likelihood, we want want to achieve $\nabla l(w) = 0$. We approximate $l(w)$ with Taylor's expansion of gradient. Namely, at iteration t, we have the update rule:

$$\nabla l(w_{t+1}) \approx \nabla l(w_t) + \nabla^2 l(w_t)(w_{t+1} - w_t)$$

To find the w such that $\nabla l(w) = 0$, we set the approximation to 0:

$$
\begin{aligned}
&0 \approx \nabla l(w_t) + \nabla^2 l(w_t)(w_{t+1} - w_t) \\
&\text{Let } g = \nabla l(w_t), \;\; H = \nabla^2 l(w_t), \;\; \Delta = (w_{t+1} - w_t) \\
&\Rightarrow -g \approx H\Delta \\
&\Rightarrow -H^{-1}g \approx \Delta
\end{aligned}
$$

Therefore, we can derive the Newton's Method update rule with our derived Hessian Matrix:

$$
\begin{aligned}
w_{t+1} &= w_t - H^{-1}g \\
&= w_t - (-XDX^\intercal)^{-1}\nabla l(w_t) \\
&= w_t + (XDX^\intercal)^{-1}\nabla l(w_t)
\end{aligned}
$$

4. Programming - Optimization. (25 pt)

This assignment guides you through building, refining, and optimizing a Logistic Regression model from scratch. Complete the `TODO` sections in the provided Python code. You can either use the Jupyter Notebook (hw2_q4.ipynb) or the Python file (hw2_q4.py).

Submission Requirements: If you use the Jupyter notebook, please make sure to convert it to hw2_q4.py and submit it to Gradescope along with hw2_q2.py and hw2_utils.py. Submit all generated plots and write the discussion sections (4(a)iii, 4(b)ii, 4(d)ii) with your written assignment.

(a) Logistic Regression

i. Start by implementing the core components of a logistic regression model. (4 pt)

A. Implement the **sigmoid function**.

B. Implement the **cost function** (binary cross-entropy). A cost function measures the total error between a model's predictions and the actual labels, summarizing its performance into a single number to be minimized. Binary cross-entropy (or Log Loss) is a specific cost function for binary classification that heavily penalizes predictions that are both confident and incorrect. Please see code for more information on how to implement this.

C. Compute the **gradients** for the weights and bias.

D. Implement the **parameter update rule** for gradient descent.

ii. **Feature Transformation:** Create new features from the existing ones using non-linear transforms (e.g., x_1^2, x_2^2, x_1x_2) to map the data into a higher-dimensional space. (4 pt)

iii. **Discussion:** Analyze the generated plot of the decision boundary and explain why the model with transformed features can now correctly separate the data, whereas a linear model in the original feature space could not. (2 pt)

(b) L1 vs. L2 Regularization

i. For the same model, modify the cost function and gradient update steps to include options for L1 and L2 regularization. Note that L1 regularization (Lasso) adds a penalty equal to the sum of the absolute values of the weights. In contrast, L2 regularization (Ridge) adds a penalty equal to the sum of the squares of the weights. (4 pt)

ii. **Discussion:** Train three separate models: one with L1 regularization, one with L2 regularization, and one with no regularization. Generate a plot that compares their decision boundaries. Compare the effects of L1 and L2 regularization on the model. What kind of effect does regularization have on the model weights? What is the main difference in how they affect the model's weights? When would you choose one over the other? (2 pt)

(c) Hyperparameter Tuning (2 pt)

i. This exercise will guide you through performing a grid search to tune the learning rate (`alpha`) and the L2 regularization (`lambda`) hyperparameters.

A. Define a set of `alpha` values and `lambda` values to test.

B. Train a model for each combination of `alpha` and `lambda`, and evaluate its performance on the validation set.

ii. Identify the best-performing hyperparameter combination. Train a final model on the full training set using these optimal values and generate a plot of its decision boundary.

(d) Gradient Descent Variants

i. Run through the implementation of Batch Gradient Descent, Mini-batch Gradient Descent, and Stochastic Gradient Descent. (6 pt)

A. **Batch Gradient Descent** uses the entire dataset to compute gradients in each step.

B. **Mini-batch Gradient Descent** processes data in small batches (e.g., size 32). In each step, you will compute the gradient and update the parameters based on just one mini-batch.

C. **Stochastic Gradient Descent (SGD)** updates the parameters after processing each single training example (i.e., a mini-batch of size 1).

ii. **Discussion:** Analyze the generated plot that visualizes the cost function over the number of updates for all three gradient descent variants. Discuss the trade-offs between the three methods. Compare them in terms of computational efficiency (speed) and stability of convergence. (1 pt)

(a) iii. Given that the input data are in circular distribution which is linearly inseparable, the Logistic Regression model, a linear model, cannot classify the input data with a linear decision boundary (shown in Figure 1).

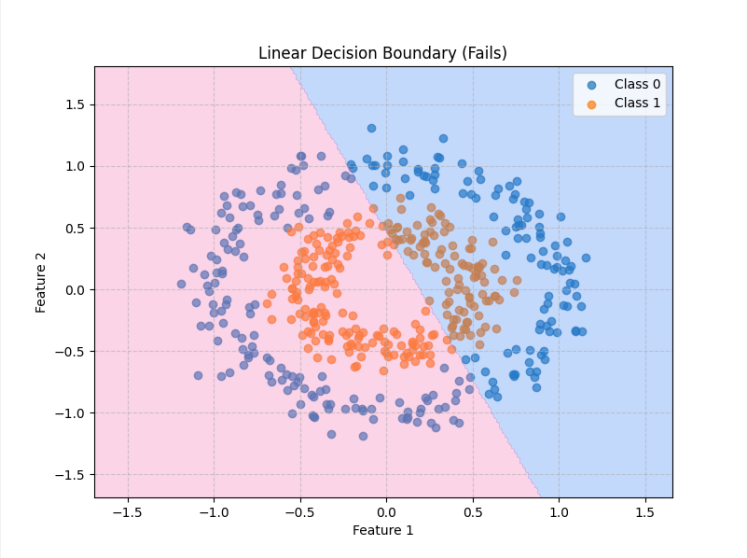

Figure 1: Linear Decision Boundary

However, by mapping the existing data points to higher dimensions, we enable our linear model to capture the non-linear patterns in the data (shown in Figure 2).

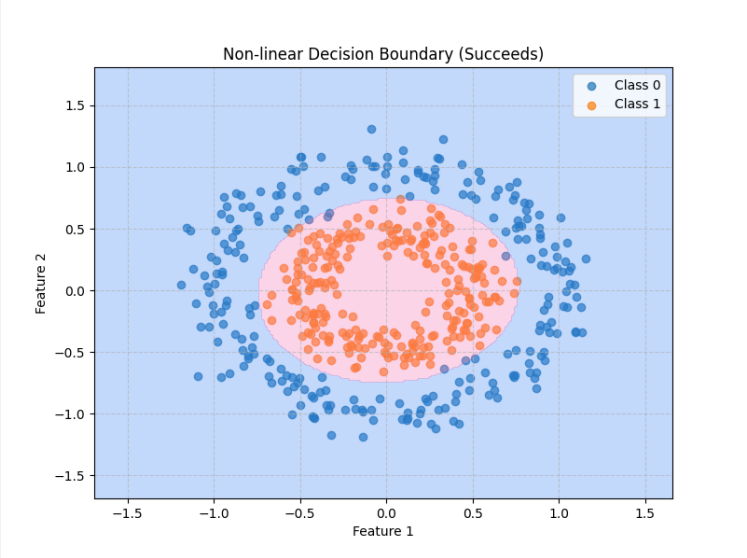

Figure 2: Non-linear Decision Boundary

(b) ii. As shown in Figure 3, both L1 and L2 regularizations bring down the weights when they are large, which suppresses the overfitting problem for the model. Both regularizations penalize large weights by introducing penalty terms ($L1 : \frac{\lambda}{m} \mid\mid w \mid\mid_1, \quad L2 : \frac{\lambda}{2m} \mid\mid w \mid\mid_2^2$). The difference is that while L2 regularization shrinks weights proportially to their current values (gradient penalty: λw) during gradient descent, the L1 regularization pulls the insignificant features to zero (gradient penalty: λ). Therefore, L1 regularization should be adopted where only certain features matter; whereas, L2 regularization should be considered when all features need to be preserved.

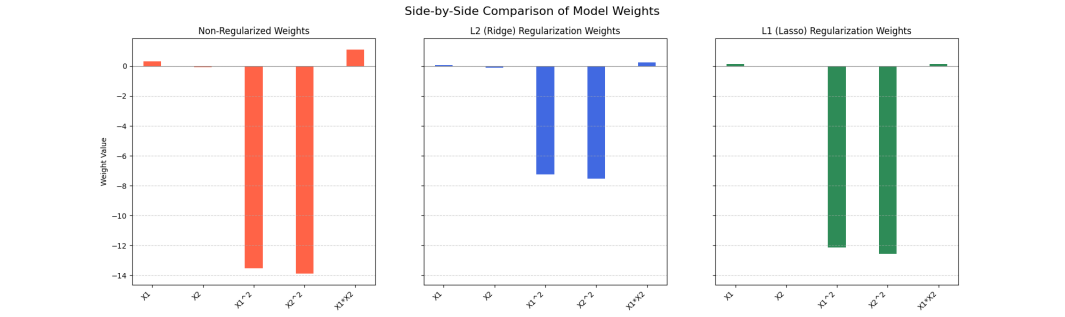

Figure 3: Comparision of Regularization Methods

(c) ii. Best Performing Hyperparameter Combination (shown in Figure 4):
— Grid Search Results —
Best parameters found: {'learning_rate': 0.1, 'lambda_val': 0.0001}
Best validation accuracy: 100.00%
Test Set Accuracy: 98.67%

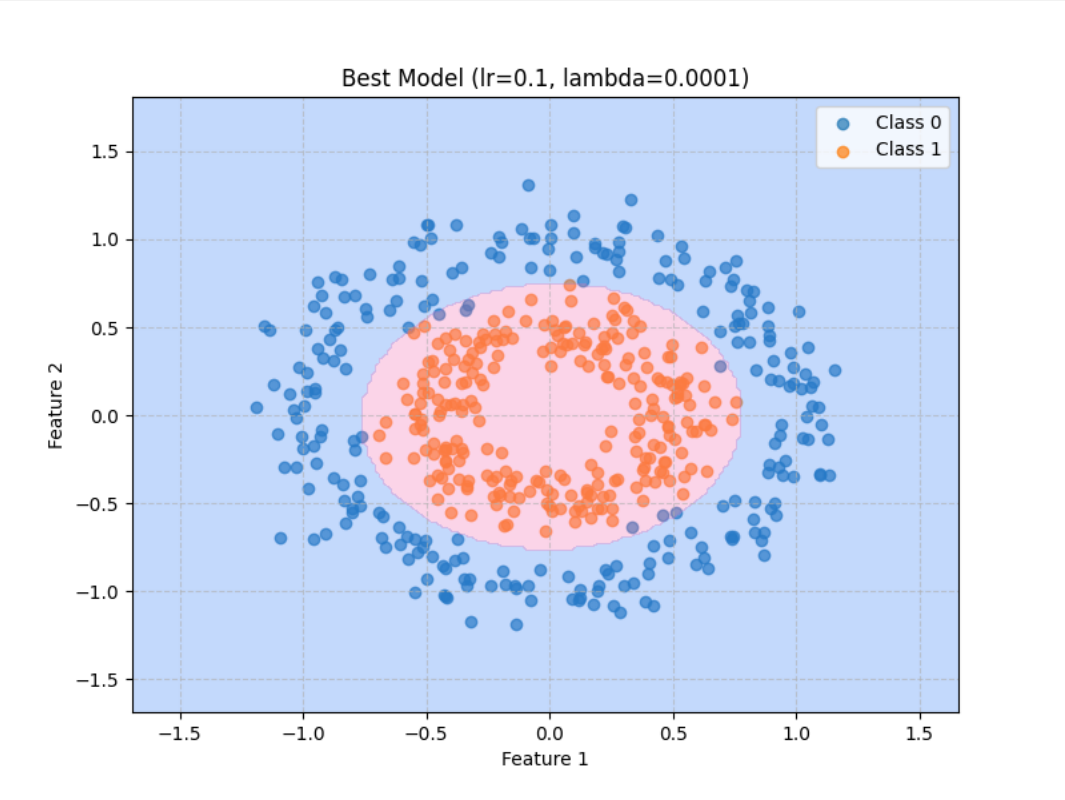

Figure 4: Best Performing Hyperparameter Combination

(d) ii. As shown in Figure 5, Stochastic GD converges the fastest while Batch GD provide the most stable and predictable convergence. Since the plot displays the result averaged by epochs, it is unable to see the oscillating individual updates of Mini-batch GD and Stochastic GD. Based on this plot, we can conclude that in terms of convergence speed: Stochastic GD >Mini-batch GD >Batch GD, as they make increasingly more parameter updates per epoch. However, in terms of stability of convergence: Batch GD >Mini-batch GD >Stochastic GD.

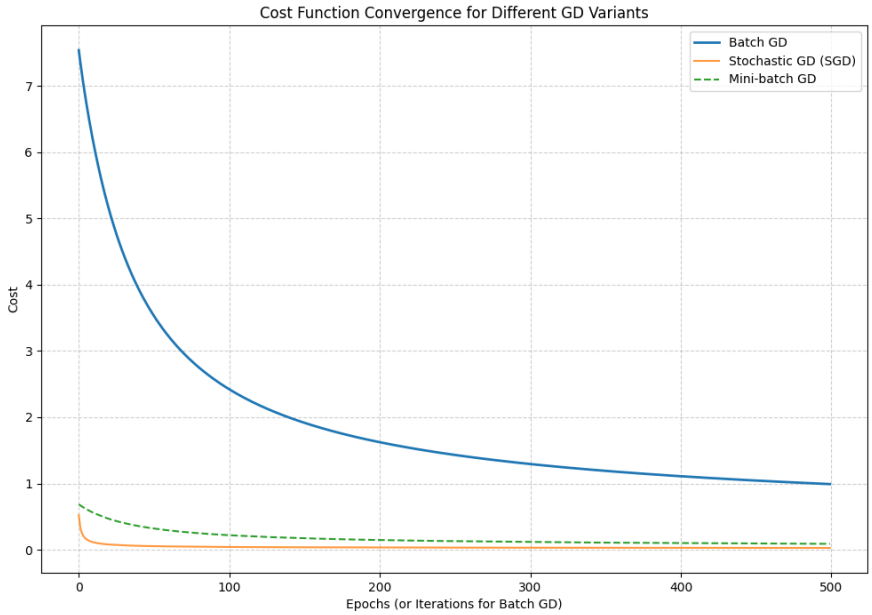

Figure 5: Comparision of Different GD Variants