Week 3 Practice

CSCI 567 Machine Learning

Spring 2025

Instructor: Haipeng Luo

1. MULTIPLE-CHOICE QUESTIONS: One or more correct choice(s) for each question.

1.1. Which of the following surrogate losses is not an upper bound of the 0-1 loss?

$\ell_{0-1}(z) = \mathbb{I}[z<0]$

(a) exponential loss: $\exp(-z)$

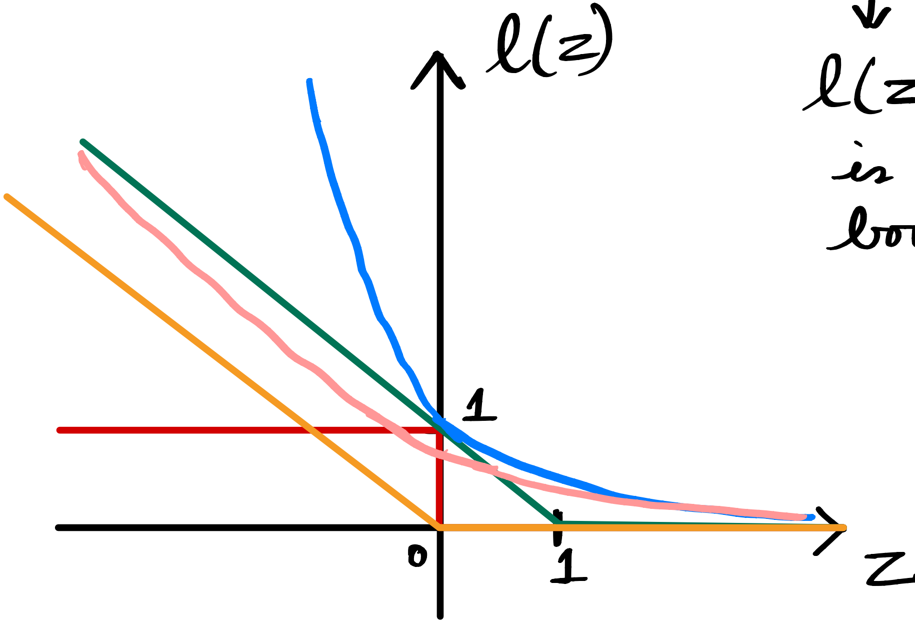
(b) hinge loss: $\max\{0, 1-z\}$

(c) perceptron loss: $\max\{0, -z\}$

(d) logistic loss: $\ln(1+\exp(-z))$

Ans: c, d. Note that here the logistic loss is using e as the base, instead of 2.

$\ell(z) = \log_2(1+e^{-z})$ is an upper bound to $\ell_{0-1}(z)$



1.2. The perceptron algorithm makes an update $\boldsymbol{w}' \leftarrow \boldsymbol{w} + \eta y_n \boldsymbol{x}_n$ with $\eta = 1$ when $\boldsymbol{w}$ misclassifies $\boldsymbol{x}_n$. Using which of the following different values for η will make sure $\boldsymbol{w}'$ classifies $\boldsymbol{x}_n$ correctly?

(a) $\eta > \frac{y(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n)}{\|\boldsymbol{x}_n\|_2^2}$ (b) $\eta < \frac{-y(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n)}{\|\boldsymbol{x}_n\|_2^2+1}$

(c) $\eta < \frac{-y(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n)}{\|\boldsymbol{x}_n\|_2^2}$ (d) $\eta > \frac{-y(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n)}{\|\boldsymbol{x}_n\|_2^2}$

Ans: d.

$$y_n = \text{sign}(w'^T x_n)$$

$$\Rightarrow y_n w'^T x_n > 0$$

$$\Rightarrow y_n (w + \eta y_n x_n)^T x_n > 0$$

$$\Rightarrow y_n w^T x_n + \eta y_n^2 \|x_n\|_2^2 > 0$$

$$\Rightarrow \eta > \frac{-y_n w^T x_n}{\|x_n\|^2}.$$

$$y_n = \pm 1$$

$$\Rightarrow y_n^2 = 1$$

1.3. Which of the following is true?

(a) Normalizing the output $\boldsymbol{w}$ of the perceptron algorithm so that $\|\boldsymbol{w}\|_2 = 1$ changes its test error.

(b) Normalizing the output $\boldsymbol{w}$ of the perceptron algorithm so that $\|\boldsymbol{w}\|_1 = 1$ changes its test error.

(c) When the data is linearly separable, logistic loss (without regularization) does not admit a minimizer.

(d) Minimizing 0-1 loss is generally NP-hard.

Ans: c, d. For c, note that when the data is separable, one can find $\boldsymbol{w}$ such that $y_n \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_n \geq 0$ for all n. Scaling this $\boldsymbol{w}$ up will always lead to smaller logistic loss $\sum_{n=1} \ln(1 + \exp(-y_n \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_n))$ and thus the function does not admit a minimizer.

(a) $w' = \dfrac{w}{\|w\|_2}$.

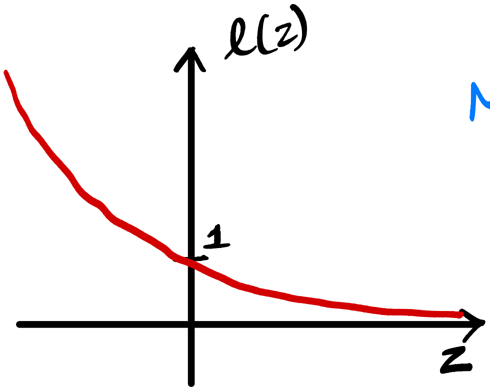
$$\hat{y}' = \text{sign}(w'^T x) = \text{sign}\left(\frac{w^T x}{\|w\|_2}\right) = \text{sign}(w^T x) = \hat{y}.$$

Prediction remains same.

(b) Same as (a).

(c) Logistic loss : $\ell(z) = \log(1 + \exp(-z))$,

where $z = y_n w^T x_n$.



Monotonically decreasing in z

Data is linearly separable

$\Rightarrow \exists\, w$ such that $y_n w^T x_n \geq 0 \;\forall n$

$\forall n,\; \ell(y_n (cw)^T x_n) \leq \ell(y_n w^T x_n)$ when $c \geq 1$

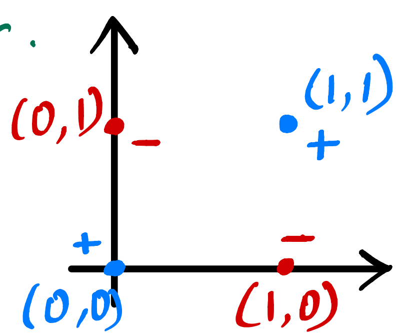
As we scale up w, loss keeps decreasing

i.e. as $c \to \infty$, $\ell \to 0$.

Thus, logistic loss does not admit minimizer.

✻ If the data is not linearly separable, minimizer may exist.

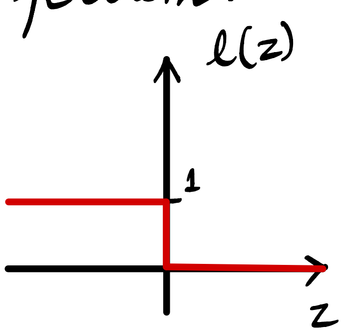
E.g.



This example is not linearly separable. Logistic loss is minimized at $w = (0,0)$.

Verify!

(d) Minimizing 0-1 loss is a non-convex, discrete optimization problem.



NP-Hard even for linear classifiers

1.4. Which of the following statement is correct for function $f(\boldsymbol{w}) = w_1 w_2$?

(a) $(0, 0)$ is the only stationary point. $\rightarrow \nabla f(w) = 0$

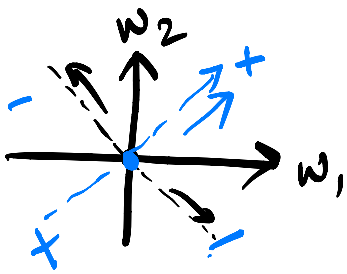
(b) $(0, 0)$ is a local minimizer.

(c) $(0, 0)$ is a local maximizer.

(d) $(0, 0)$ is a saddle point.

Ans: a, d. The gradient is $\nabla f(\boldsymbol{w}) = (w_2, w_1)$, so the only stationary point is $(0, 0)$.

$$\nabla f(w) = \begin{pmatrix} w_2 \\ w_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow w_1 = w_2 = 0 .$$



Consider $w_2 = -w_1$:

$f(w) = -w_1^2 < 0 = f((0,0))$

$(0,0)$ is not a local minimizer.

Consider $w_2 = w_1$:

$f(w) = w_1^2 > 0 = f((0,0))$

$(0, 0)$ is not a local maximizer.

2. Perceptron

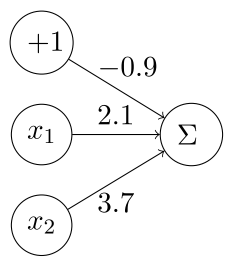
Consider the following training dataset:

x	y
(0, 0)	-1
(0, 1)	-1
(1, 0)	-1
(1, 1)	1

$\hat{y} = sign(w^T x)$

and a perceptron with weights $(w_0, w_1, w_2) = \{-0.9, 2.1, 3.7\}$

$W = \begin{pmatrix} -0.9 \\ 2.1 \\ 3.7 \end{pmatrix}$



$x = \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}$

2.1. What is the accuracy of the perceptron on the training data?

SOLUTION:

x	y	$\hat{y}$	$\mathbb{I}(y = \hat{y})$
(0, 0)	-1	$sgn(-0.9) = -1$	Y
(0, 1)	-1	$sgn(-0.9 + 3.7) = 1$	N
(1, 0)	-1	$sgn(-0.9 + 2.1) = 1$	N
(1, 1)	1	$sgn(-0.9 + 2.1 + 3.7) = 1$	Y

Out of four predictions, two are correct. The accuracy is hence 50%.

3^{rd} data point

2.2. Select $\mathbf{x} = (1, 0)$ and $y = -1$. Use the perceptron training rule with $\eta = 1$ to train the perceptron for one iteration. What are the weights after this iteration?

For the given $\mathbf{x} = (1, 0)$ the classifier makes a mistake ($\hat{y} = 1$). We need to update the weights following the perceptron rule.

$$w' \leftarrow \underline{w + \eta y x} = \begin{pmatrix} -0.9 \\ 2.1 \\ 3.7 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1.9 \\ 1.1 \\ 3.7 \end{pmatrix}$$

2.3. What is the accuracy of the perceptron on the training data after this iteration? Does the accuracy improve?

SOLUTION:

$\mathbf{x}$	y	$\hat{y}$	$\mathbb{I}(y = \hat{y})$
(0, 0)	-1	$sgn(-1.9) = -1$	Y
(0, 1)	-1	$sgn(-1.9 + 3.7) = 1$	N
(1, 0)	-1	$sgn(-1.9 + 1.1) = -1$	Y
(1, 1)	1	$sgn(-1.9 + 1.1 + 3.7) = 1$	Y

classified correctly

With the new weights, three out of four are correct, hence accuracy increased to 75%.

3. Maximum Likelihood Estimation

A random sample set $X_1, X_2, \dots, X_n$ of size n is taken from a Poisson distribution with a mean of $\lambda > 0$. As a reminder, a Poisson distribution is a discrete probability distribution over the natural numbers, with the following probability mass function

$$P(X = x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}, \ \forall x \in \{0, 1, 2, \dots, \}$$

3.1. Find the log likelihood of the data; call it $l(\lambda)$. You may use any log base you want.

likelihood of the data $= \prod_{i=1}^{n} P(X = X_i)$

log-likelihood $= \ell(\lambda) = \log \prod_{i=1}^{n} P(X = X_i)$

$$= \sum_{i=1}^{n} \log\left(\frac{\lambda^{X_i} e^{-\lambda}}{X_i!}\right)$$

$$= \sum_{i=1}^{n} X_i \log\lambda - \lambda - \log(X_i!)$$

$$= \log\lambda \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log(X_i!)$$

3.2. Find the maximum likelihood estimator for λ.

Maximize $\ell(\lambda)$

$$\ell'(\lambda) = 0 \Rightarrow \frac{1}{\lambda}\sum_{i=1}^{n} X_i - n = 0$$

$$\Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} X_i}{n}$$ average

$$\ell''(\hat{\lambda}) = -\frac{1}{\hat{\lambda}^2}\sum_{i=1}^{n} X_i < 0 \Rightarrow \hat{\lambda}$$ is a maximizer