CSCI 567 Discussion Section
Week 3

Problem 1

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?

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

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

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

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

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

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.

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.

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

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

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

Problem 1.1

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

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

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

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

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

Problem 1.1

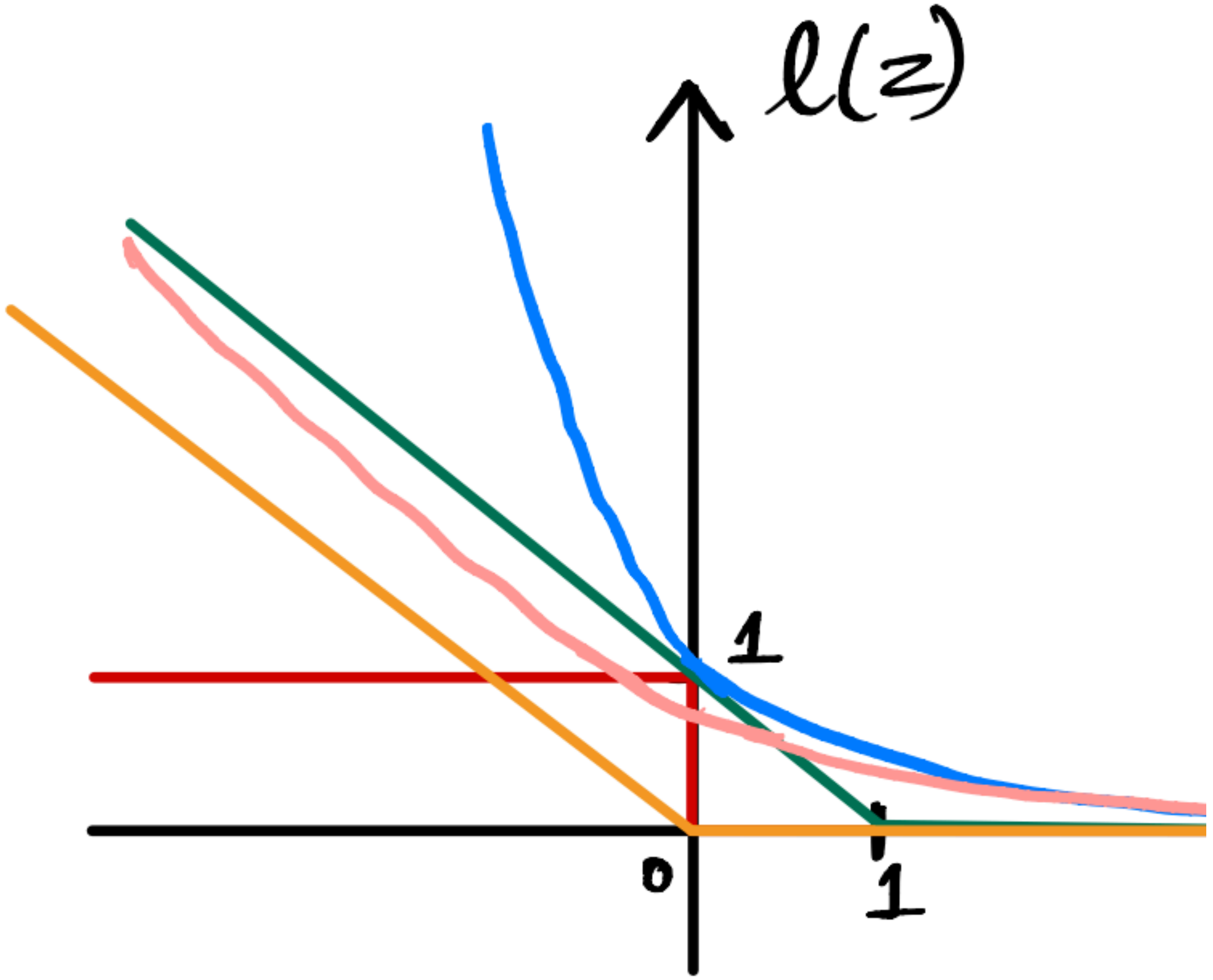
1.1. Which of the following surrogate losses is not an upper bound of the 0-1 loss?

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

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

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

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



Problem 1.2

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

Problem 1.2

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

$$
\begin{aligned}
y_n &= \mathrm{sgn}(\mathbf{w}'^T \mathbf{x}_n) \\
&\Rightarrow y_n \mathbf{w}'^T \mathbf{x}_n > 0 \\
&\Rightarrow y_n (\mathbf{w} + \eta y_n \mathbf{x}_n)^T \mathbf{x}_n > 0 \\
&\Rightarrow y_n \mathbf{w}^T \mathbf{x}_n + \eta y_n^2 \|\mathbf{x}_n\|_2^2 > 0 \\
&\Rightarrow \eta > \frac{-y_n \mathbf{w}^T \mathbf{x}_n}{\|\mathbf{x}_n\|_2^2}
\end{aligned}
$$

Problem 1.3

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.

Problem 1.3

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.

Problem 1.4

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.
(b) $(0, 0)$ is a local minimizer.
(c) $(0, 0)$ is a local maximizer.
(d) $(0, 0)$ is a saddle point.

Problem 1.4

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.
(b) $(0,0)$ is a local minimizer.
(c) $(0,0)$ is a local maximizer.
(d) $(0,0)$ is a saddle point.

The gradient of $f(w)$ is:

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

Case 1: Consider $w_2 = -w_1$

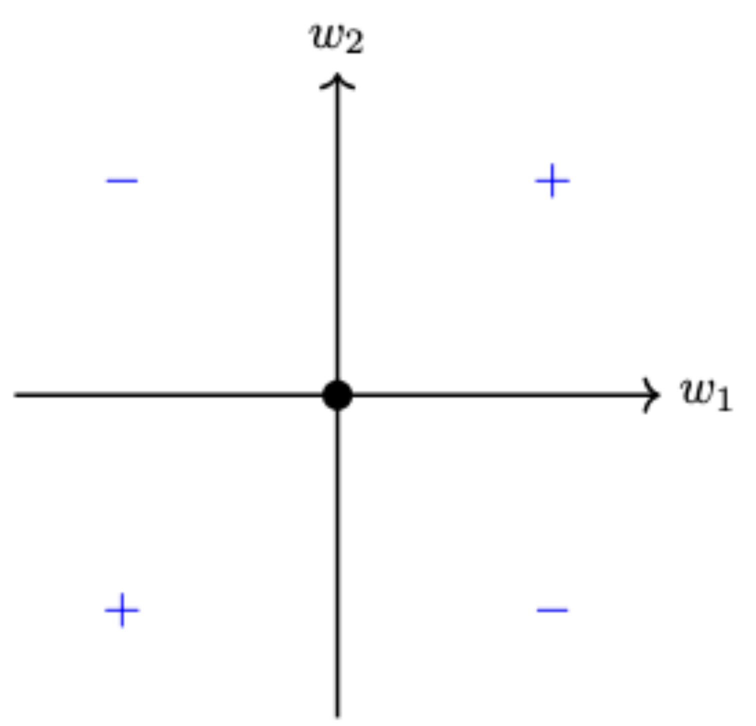
$$f(w) = -w_1^2 < 0 = f((0,0)) \tag{1}$$
$$\tag{2}$$

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

Case 2: Consider $w_2 = w_1$

$$f(w) = w_1^2 > 0 = f((0,0)) \tag{3}$$

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



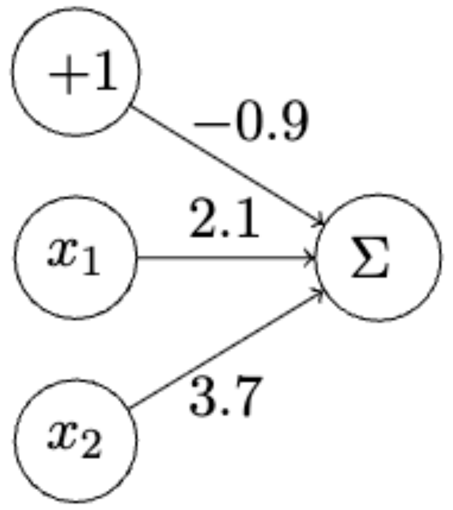
Problem 2

2. Perceptron

Consider the following training dataset:

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

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



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

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?

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

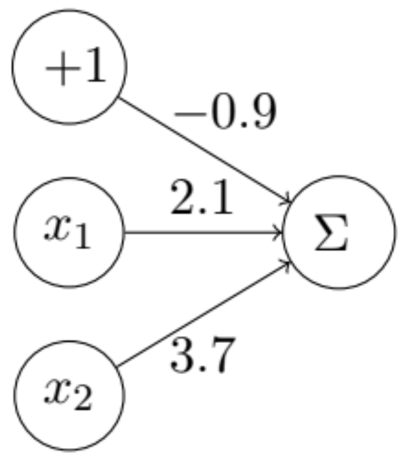
Problem 2.1

2. Perceptron

Consider the following training dataset:

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

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



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

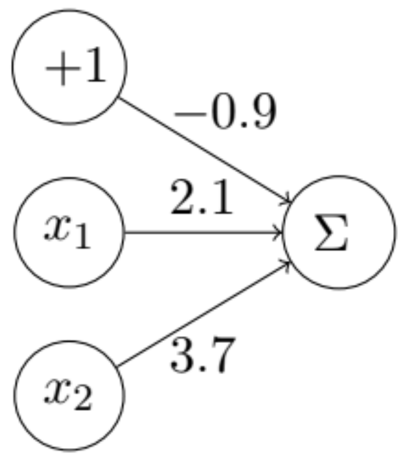
Problem 2.1

2. Perceptron

Consider the following training dataset:

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

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



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

Accuracy = 50%

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

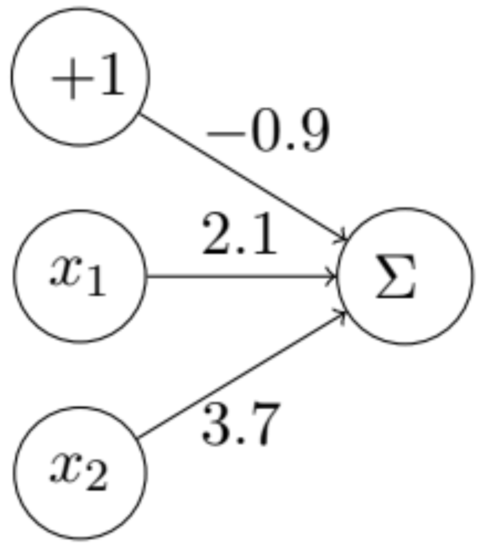
Problem 2.2

2. Perceptron

Consider the following training dataset:

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

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



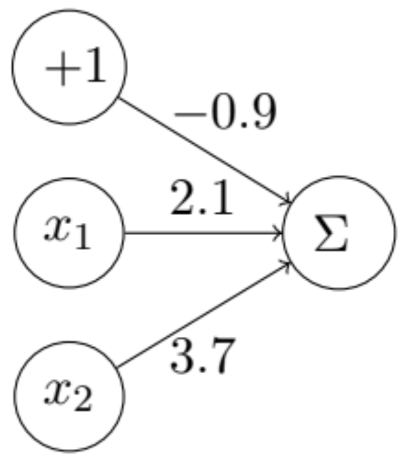
2.2. Select $\mathbf{x} = (1, 0)$ and $y = -1$. Use the perceptron training rule with $\eta = 1$ to train the perceptron for one iteration. What are the weights after this iteration?

Problem 2.2

2. Perceptron
Consider the following training dataset:

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

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



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.

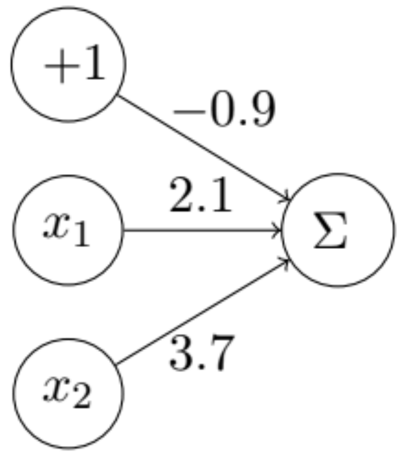
$$\mathbf{w}' \leftarrow \mathbf{w} + \eta y \mathbf{x} = \begin{pmatrix} -0.9 \\ 2.1 \\ 3.7 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1.9 \\ 1.1 \\ 3.7 \end{pmatrix}$$

Problem 2.3

2. Perceptron
Consider the following training dataset:

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

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



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

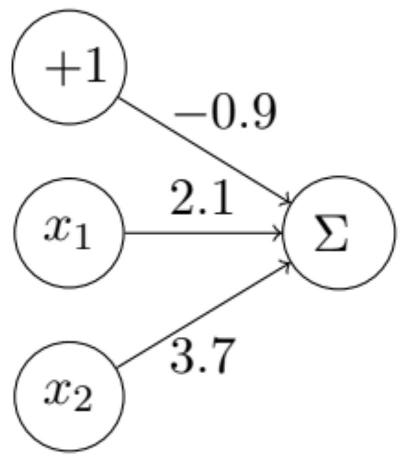
Problem 2.3

2. Perceptron

Consider the following training dataset:

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

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



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

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

Accuracy = 75%

Problem 3

3. Maximum Likelihood Estimation

A random sample set $X_1, X_2, \ldots, X_n$ 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

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

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

3.2. Find the maximum likelihood estimator for λ.

Problem 3.1

3. Maximum Likelihood Estimation

A random sample set $X_1, X_2, \ldots, X_n$ 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

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

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

Problem 3.1

3. Maximum Likelihood Estimation

A random sample set $X_1, X_2, \ldots, X_n$ 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

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

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

$$\text{Likelihood of the data} = \prod_{i=1}^{n} P(X = x_i)$$

$$\begin{aligned}
\text{log-likelihood} = \ell(\lambda) &= \log \prod_{i=1}^{n} P(X = x_i) \\
&= \sum_{i=1}^{n} \log\left(\frac{\lambda^{X_i} e^{-\lambda}}{x_i!}\right) \\
&= \sum_{i=1}^{n} x_i \log\lambda - \lambda - \log(x_i!) \\
&= \log\lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log(x_i!)
\end{aligned}$$

Problem 3.2

3. Maximum Likelihood Estimation

A random sample set $X_1, X_2, \ldots, X_n$ 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

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

3.2. Find the maximum likelihood estimator for λ.

Problem 3.2

3. Maximum Likelihood Estimation

A random sample set $X_1, X_2, \ldots, X_n$ 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

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

3.2. Find the maximum likelihood estimator for λ.

Maximize $\ell(\lambda)$

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

$$\Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \leftarrow \text{average}$$

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