Discussion 6

Note: *Your TA will probably not cover all the problems on this worksheet. The discussion worksheets are not designed to be finished within an hour. They are deliberately made slightly longer so they can serve as resources you can use to practice, reinforce, and build upon concepts discussed in lectures, discussions, and homework.*

This Week's Cool AI Demo/Video:
Robot dog: `https://www.youtube.com/watch?v=iI8UUu9g8iI`

Super fast recovery on humanoids: `https://www.youtube.com/watch?v=bPSLMX_V38E`

Ping pong with humanoids: `https://www.youtube.com/watch?v=tOfPKW6D3gE`

1 Evaluation Metrics and Threshold Selection

You trained a logistic regression model that outputs a probability $\hat{p}(y{=}1 \mid x)$ for each test example. The following table shows the true labels and predicted probabilities for a small test set:

ID	**True** y	$\hat{p}$
1	1	0.93
2	1	0.84
3	0	0.72
4	0	0.63
5	0	0.58
6	1	0.49
7	0	0.41
8	0	0.35
9	1	0.32
10	0	0.18

There are 4 positive and 6 negative examples. The model predicts $\hat{y} = 1$ when $\hat{p} \geq \tau$, where $\tau \in (0, 1)$ is the binary classification threshold. Recall that for binary classification, a prediction can either by a true positive (TP), true negative (TN), false positive (FP), or false negative (FN).

For this problem, we will consider three metrics to evaluate classification performance:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

(a) Explain in words the difference between accuracy, precision, and recall for binary classification.

Solution:

Accuracy measures the overall fraction of predictions that are correct, including both positives and negatives.

Precision measures the fraction of predicted positives that are actually positive. It answers: "Of all the times the model predicted 'positive', how often was it correct?"

Recall measures the fraction of actual positives that were correctly predicted. It answers: "Of all the actual positives, how many did the model correctly identify?"

(b) For each threshold $\tau \in \{0.3, 0.5, 0.8\}$, compute the elements of the binary confusion matrix: TP, FP, TN, FN. Then use these values to calculate accuracy, precision, and recall.

Solution:

(a) $\tau = 0.30$: predicted positives = IDs 1–9

- TP = 4, FP = 5, TN = 1, FN = 0

$$\text{Accuracy} = 0.50, \quad \text{Precision} = \frac{4}{9} \approx 0.44, \quad \text{Recall} = 1.00,$$

(b) At $\tau = 0.50$: predicted positives = IDs 1–5.

- TP = 2, FP = 3, TN = 3, FN = 2

$$\text{Accuracy} = 0.50, \quad \text{Precision} = 0.40, \quad \text{Recall} = 0.50,$$

(c) $\tau = 0.80$: predicted positives = IDs 1-2

- TP = 2, FP = 0, TN = 6, FN = 2

$$\text{Accuracy} = 0.80, \quad \text{Precision} = 1.00, \quad \text{Recall} = 0.50,$$

(c) How would you choose the best threshold for this dataset?

i. Consider a **disease screening** scenario, where $y = 1$ indicates that a patient actually has the disease and $\hat{y} = 1$ indicates that we detect the disease and provide treatment. Which threshold from part (b) would you choose? Provide a brief justification for your answer.

ii. How does your answer change if the context is **spam detection**, where $y = 1$ indicates an email is spam and $\hat{y} = 1$ indicates that we detect spam and immediately delete it?

Solution:

i. For disease screening, we prioritize **high recall** because we prefer false positives (treating a healthy person) over false negatives (overlooking a person with the disease).

From part (b):

- $\tau = 0.30$: Recall = **1.00**
- $\tau = 0.50$: Recall = 0.50
- $\tau = 0.80$: Recall = 0.50

Thus, choose $\boxed{\tau = 0.30}$. This ensures that we detect more patients with the disease.

ii. For spam detection, we prioritize **high precision** because we prefer false negatives (letting a spam email through) over false positives (deleting a good email).

From part (b):

- $\tau = 0.30$: Precision = 0.44
- $\tau = 0.50$: Precision = 0.4
- $\tau = 0.80$: Precision = **1.00**

Thus, choose $\boxed{\tau = 0.80}$. This avoids deleting legitimate emails while catching some spam.

2 Statistical Justification of Logistic Regression

Assume that we have N i.i.d. data points $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$, where each y_n is a binary label in $\{0, 1\}$. We model the posterior probability of the labels given the observed features as a Bernoulli distribution, where the probability of a positive sample is given by the sigmoid function, meaning

$$p(Y = y|\mathbf{x}; \mathbf{w}) = p^y(1-p)^{1-y}, \quad \text{where } p = \sigma(\mathbf{w}^\top \mathbf{x}) \text{ and } \sigma(a) = \frac{1}{1+e^{-a}}.$$

(a) Show that for a given data point $\mathbf{x}$, the log ratio of the conditional probabilities, or *log odds*, is linear in $\mathbf{x}$. More specifically, show that

$$\log \frac{p(Y = 1 \mid \mathbf{x}; \mathbf{w})}{p(Y = 0 \mid \mathbf{x}; \mathbf{w})} = \mathbf{w}^\top \mathbf{x}.$$

Solution:

$$\begin{aligned}
\log \frac{p(Y = 1 \mid \mathbf{x}; \mathbf{w})}{p(Y = 0 \mid \mathbf{x}; \mathbf{w})} &= \log \frac{p}{1-p} \\
&= \log \frac{\frac{1}{1+e^{-\mathbf{w}^\top \mathbf{x}}}}{\frac{e^{-\mathbf{w}^\top \mathbf{x}}}{1+e^{-\mathbf{w}^\top \mathbf{x}}}} \\
&= \log \frac{1}{e^{-\mathbf{w}^\top \mathbf{x}}} \\
&= \mathbf{w}^\top \mathbf{x}
\end{aligned}$$

(b) Starting from the Bernoulli likelihood, derive the logistic loss using Maximum Likelihood Estimation (MLE). Show that maximizing the likelihood of a Bernoulli model is equivalent to minimizing the logistic loss.

Solution:

Step 1: Bernoulli PMF

For a Bernoulli random variable Y with parameter $p_n = \sigma(\mathbf{w}^\top \mathbf{x}_n)$,

$$p(Y = y_n; \mathbf{w}) = p_n^{y_n}(1-p_n)^{1-y_n}.$$

Step 2: Likelihood

The likelihood of a dataset $\mathcal{D}$ with N i.i.d. observations is then

$$p(\mathcal{D}|\mathbf{w}) = \prod_{n=1}^N p(Y = y_n; \mathbf{w}) = \prod_{n=1}^N p_n^{y_n}(1-p_n)^{1-y_n}.$$

Step 3: Log-likelihood

Taking the natural log of our likelihood, we have

$$\log p(\mathcal{D}|\mathbf{w}) = \sum_{n=1}^N [y_n \log p_n + (1-y_n)\log(1-p_n)].$$

Step 4: Negative log-likelihood

The negative log likelihood of our dataset is then

$$-\log p(\mathcal{D}|\mathbf{w}) = -\sum_{n=1}^{N}[y_n \log p_n + (1-y_n)\log(1-p_n)].$$

Step 5: Logistic Loss

Maximizing the likelihood is equivalent to minimizing the negative log likelihood:

$$\underset{\mathbf{w}}{\text{argmax}}\ p(\mathcal{D}|\mathbf{w}) \equiv \underset{\mathbf{w}}{\text{argmin}}\ -\sum_{n=1}^{N}[y_n \log p_n + (1-y_n)\log(1-p_n)].$$

Conclusion: The logistic loss function is the negative log-likelihood of the Bernoulli model. Thus, minimizing logistic loss is equivalent to maximizing the likelihood under the assumption that labels follow independent Bernoulli distributions.