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.

(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.

(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?

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

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

(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.