COMP 584 — Homework 1

Dev Sanghvi
NetID: ds221

February 8, 2026

Problem 1 (20 pts): Text Representation (BoW and TF-IDF)

Vocabulary (fixed order):

$$\mathcal{V} = [\text{lecture}, \text{homework}, \text{exam}, \text{student}, \text{grade}, \langle\text{unk}\rangle]$$

Documents:
- d_1: lecture homework homework quiz
- d_2: exam grade grade student
- d_3: student lecture exam

Tokenization rule: Split on whitespace. Any token not in $\mathcal{V}$ is mapped to $\langle\text{unk}\rangle$.

Part I (12 pts): Bag-of-Words (Binary vs. Count)

1. (4 pts) Binary BoW vector for document d_1:
Tokens in d_1: [lecture, homework, homework, quiz]
Mapping:
- "lecture" $\in \mathcal{V} \Rightarrow$ index 0
- "homework" $\in \mathcal{V} \Rightarrow$ index 1
- "quiz" $\notin \mathcal{V} \Rightarrow$ maps to $\langle\text{unk}\rangle$ at index 5

Binary BoW (1 if token appears at least once, 0 otherwise):

$$\boxed{\mathbf{x}_{d_1}^{\text{binary}} = [1, 1, 0, 0, 0, 1]}$$

2. (4 pts) Count BoW vector for document d_1:
Token counts:
- lecture: 1
- homework: 2
- quiz $\rightarrow \langle\text{unk}\rangle$: 1

$$\boxed{\mathbf{x}_{d_1}^{\text{count}} = [1, 2, 0, 0, 0, 1]}$$

3. (4 pts) Count BoW vector for document d_2:
Tokens in d_2: [exam, grade, grade, student]
Token counts:
- exam: 1

- grade: 2
- student: 1

$$\boxed{\mathbf{x}_{d_2}^{\text{count}} = [0, 0, 1, 1, 2, 0]}$$

Part II (8 pts): TF-IDF

$N = 3$ documents in the corpus.

1. (4 pts) Compute $\mathbf{df}(t)$ for each token $t \in \mathcal{V}$:

Token	**Appears in**	$\mathbf{df}(t)$
lecture	d_1, d_3	2
homework	d_1	1
exam	d_2, d_3	2
student	d_2, d_3	2
grade	d_2	1
$\langle$unk$\rangle$	d_1 (from "quiz")	1

$$\boxed{\text{df} = [2, 1, 2, 2, 1, 1]}$$

2. (4 pts) TF-IDF vector for document d_1:

Recall: $\text{idf}(t) = \ln\left(\frac{N}{\text{df}(t)}\right)$ and $\text{tf-idf}(t, d) = \text{tf}(t, d) \cdot \text{idf}(t)$.

Token	$\mathbf{tf}(t, d_1)$	$\mathbf{idf}(t)$	$\mathbf{tf} \cdot \mathbf{idf}$
lecture	1	$\ln(3/2)$	$\ln(3/2)$
homework	2	$\ln(3/1) = \ln 3$	$2 \ln 3$
exam	0	$\ln(3/2)$	0
student	0	$\ln(3/2)$	0
grade	0	$\ln 3$	0
$\langle$unk$\rangle$	1	$\ln 3$	$\ln 3$

$$\boxed{\mathbf{x}_{d_1}^{\text{tf-idf}} = \left[\ln \tfrac{3}{2},\ 2 \ln 3,\ 0,\ 0,\ 0,\ \ln 3\right]}$$

Problem 2 (30 pts): Classification

Throughout this problem, consider a single training example $(\mathbf{x}, y)$ with $y \in \{0, 1\}$, and use the sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$.

Part I (12 pts): A Simple Two-Layer MLP

Model:

$$\mathbf{h} = W^{(1)}\mathbf{x}, \quad z = \mathbf{w}^{(2)\top}\mathbf{h} + b, \quad \hat{y} = \sigma(z)$$

Loss:

$$\mathcal{L} = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})]$$

1. (5 pts) Why maximize log-likelihood instead of likelihood?

(a) **Numerical Stability:** Likelihoods are products of probabilities, which can become extremely small (underflow) for large datasets. Taking the logarithm converts products into sums, preventing numerical underflow.

(b) **Convexity:** For many models (e.g., logistic regression), the log-likelihood is a concave function of the parameters, making optimization easier with gradient-based methods. The raw likelihood is not convex.

(c) **Gradient Simplicity:** Derivatives of log-likelihoods often have simpler, cleaner forms. For example, gradients involve $(y - \hat{y})$ terms rather than complex products.

2. (7 pts) Derive negative log-likelihood for $y \sim$ Bernoulli$(\hat{y})$:
The Bernoulli probability mass function is:

$$P(y \mid \hat{y}) = \hat{y}^y (1 - \hat{y})^{1-y}$$

Taking the log:

$$\log P(y \mid \hat{y}) = y \log \hat{y} + (1 - y) \log(1 - \hat{y})$$

The negative log-likelihood (NLL) is:

$$\mathcal{L} = -\log P(y \mid \hat{y}) = -\left[y \log \hat{y} + (1 - y) \log(1 - \hat{y})\right]$$

$$\boxed{\mathcal{L} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})}$$

This is exactly the **binary cross-entropy loss**.

Part II (4 pts): Forward Pass

Given model:

$$\begin{aligned}
\mathbf{a}_1 &= W_1\mathbf{x} \\
\mathbf{a}_{21} &= W_{21}\mathbf{a}_1, \quad \mathbf{a}_{22} = W_{22}\mathbf{a}_1 \\
\mathbf{a}_2 &= \mathbf{a}_{21} + \mathbf{a}_{22} \\
a_3 &= \mathbf{w}_3^\top \mathbf{a}_2, \quad a_u = \mathbf{u}^\top \mathbf{x} \\
z &= a_3 + a_u, \quad \hat{y} = \sigma(z)
\end{aligned}$$

Substituting:

$$\begin{aligned}\mathbf{a}_2 &= W_{21}W_1\mathbf{x} + W_{22}W_1\mathbf{x} = (W_{21} + W_{22})W_1\mathbf{x}\\ a_3 &= \mathbf{w}_3^\top (W_{21} + W_{22})W_1\mathbf{x}\\ z &= \mathbf{w}_3^\top (W_{21} + W_{22})W_1\mathbf{x} + \mathbf{u}^\top\mathbf{x}\end{aligned}$$

$$\boxed{\hat{y} = \sigma\left(\mathbf{w}_3^\top (W_{21} + W_{22})W_1\mathbf{x} + \mathbf{u}^\top\mathbf{x}\right)}$$

Part III (7 pts): Backward Pass

Using the chain rule with $\mathcal{L} = -[y \log \hat{y} + (1-y)\log(1-\hat{y})]$.

First, useful derivative:

$$\frac{\partial \mathcal{L}}{\partial z} = \hat{y} - y$$

Gradient w.r.t. u:

Since $z = a_3 + \mathbf{u}^\top\mathbf{x}$:

$$\frac{\partial z}{\partial \mathbf{u}} = \mathbf{x}$$

$$\boxed{\frac{\partial \mathcal{L}}{\partial \mathbf{u}} = (\hat{y} - y)\mathbf{x}}$$

Gradient w.r.t. W_1:

Chain rule through the network:

$$\frac{\partial \mathcal{L}}{\partial W_1} = \frac{\partial \mathcal{L}}{\partial z} \cdot \frac{\partial z}{\partial \mathbf{a}_2} \cdot \frac{\partial \mathbf{a}_2}{\partial \mathbf{a}_1} \cdot \frac{\partial \mathbf{a}_1}{\partial W_1}$$

Where:

- $\frac{\partial z}{\partial \mathbf{a}_2} = \mathbf{w}_3$
- $\frac{\partial \mathbf{a}_2}{\partial \mathbf{a}_1} = W_{21}^\top + W_{22}^\top = (W_{21} + W_{22})^\top$
- $\frac{\partial \mathbf{a}_1}{\partial W_1} = \mathbf{x}^\top$

$$\boxed{\frac{\partial \mathcal{L}}{\partial W_1} = (\hat{y} - y)(W_{21} + W_{22})^\top \mathbf{w}_3\mathbf{x}^\top}$$

Part IV (7 pts): Backward Pass with W_{super}

Modified model:

$$\mathbf{a}_{21} = W_{21}W_{\text{super}}\mathbf{a}_1, \quad \mathbf{a}_{22} = W_{22}W_{\text{super}}\mathbf{a}_1$$

1. Gradient w.r.t. W_{super}:

Let $\mathbf{a}_s = W_{\text{super}}\mathbf{a}_1$. Then $\mathbf{a}_2 = (W_{21} + W_{22})\mathbf{a}_s$.

$$\frac{\partial \mathcal{L}}{\partial W_{\text{super}}} = \frac{\partial \mathcal{L}}{\partial z} \cdot \frac{\partial z}{\partial \mathbf{a}_2} \cdot \frac{\partial \mathbf{a}_2}{\partial \mathbf{a}_s} \cdot \frac{\partial \mathbf{a}_s}{\partial W_{\text{super}}}$$

$$\boxed{\frac{\partial \mathcal{L}}{\partial W_{\text{super}}} = (\hat{y} - y)(W_{21} + W_{22})^\top \mathbf{w}_3 \mathbf{a}_1^\top}$$

2. Gradient w.r.t. W_1:
Now the chain extends through W_{super}:

$$\boxed{\frac{\partial \mathcal{L}}{\partial W_1} = (\hat{y} - y) W_{\text{super}}^\top (W_{21} + W_{22})^\top \mathbf{w}_3 \mathbf{x}^\top}$$