# Naïve Bayes Example

- Reference: https://web.stanford.edu/~jurafsky/slp3/, Chapter 3.

Positive tweets

I am happy because I am learning NLP

I am happy, not sad.

Negative tweets

I am sad, I am not learning NLP

I am sad, not happy

| word | Pos | Neg |
|---|---|---|
| I | 3 | 3 |
| am | 3 | 3 |
| happy | 2 | 1 |
| because | 1 | 0 |
| learning | 1 | 1 |
| NLP | 1 | 1 |
| sad | 1 | 2 |
| not | 1 | 2 |
| $N_{class}$ | 13 | 12 |

$Pr(x|Pos)$ → $Pr(x|Neg)$

| word | Pos | Neg |
|---|---|---|
| I | 0.24 | 0.25 |
| am | 0.24 | 0.25 |
| happy | 0.15 | 0.08 |
| because | 0.08 | 0 |
| learning | 0.08 | 0.08 |
| NLP | 0.08 | 0.08 |
| sad | 0.08 | 0.17 |
| not | 0.08 | 0.17 |

$Pr(Pos) = \frac{13}{25}$

$Pr(Neg) = \frac{12}{25}$

Let's classify the following tweets as **positive or negative**:

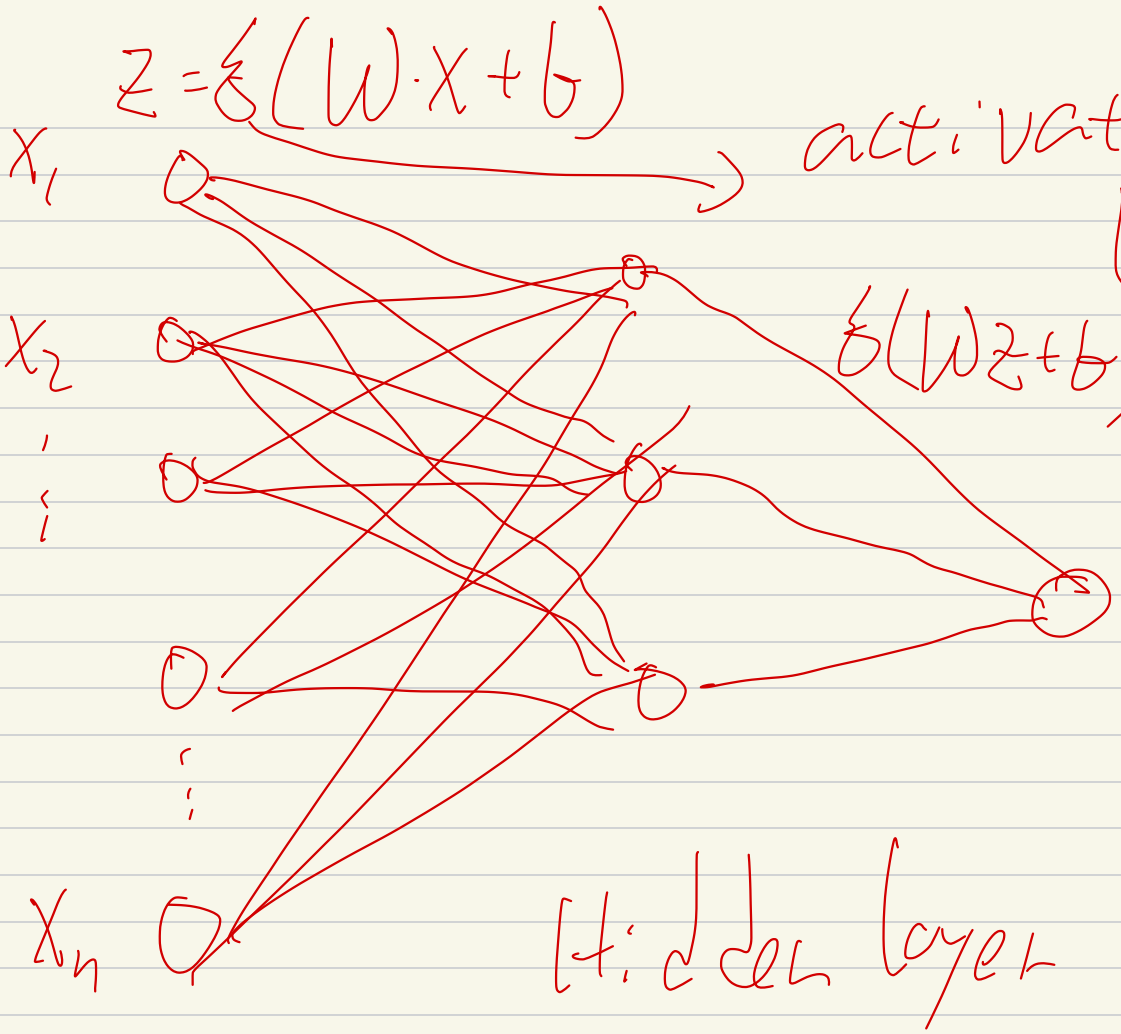
1. I am not sad.

$Pr(Pos) \cdot Pr(I|Pos) \cdot Pr(am|Pos) \cdot Pr(not|Pos) \cdot Pr(sad|Pos)$

$Pr(Neg) \cdot Pr(I|Neg) \cdot Pr(am|Neg) \cdot Pr(not|Neg) \cdot Pr(sad|Neg)$

$\frac{13}{25} \cdot \frac{0.24}{0.15} \cdot \frac{0.24}{0.5} \cdot \frac{0.08}{0.17} \cdot \frac{0.08}{0.17} < 1$

2. I am learning NLP.




$$Pr(Y=1 \mid X)$$

$$\hat{y} = \phi(W \xi(WX+b)+b)$$

① How to obtain $\theta \triangleq (W, b)$ ?

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i, \theta))$$

↳ DNN.

② What is $\mathcal{L}$? { ① regression: $(y-\hat{y})^2$
② classification
$-y \log \hat{y} - (1-y)\log(1-\hat{y})$

③ How to find $\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta)$

Gradient Descent.

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \alpha \cdot \nabla_{\theta} \mathcal{L}(\hat{\theta}_t)$$

④ How to estimate $\nabla_\theta \mathcal{L}(\hat{\theta}_t)$

(i) Chain-Rule / BP

$$y = f(z) \qquad z = g(x)$$

$$\frac{dy}{dx} = \frac{df}{dz}\bigg|_{z=g(x)} \cdot \frac{dz}{dx}\bigg|_{x}$$

(ii) SGD: Shuffle data observations

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \alpha \cdot \nabla_\theta \mathcal{L}(Y_i, f(X_i; \hat{\theta}_t))$$

# Adam  Default Choice of optimizer

(1) Initialization $(w, b) \sim N\left(0, \frac{1}{2n_l}\right)$

(2) Normalization: $\frac{x_i - \bar{x}_i}{\sigma_i}$

(3) Skip Connection

$g_1 \cdot g_2 \cdots g_n$  Vanishing Gradient

$$f(x) = \underbrace{-x + f(x)} + x$$