Recitation: Classification & Representation Learning

Evan Hernandez–Ekin Akyürek

MIT 6.864/806 – Spring 2021

Agenda

1. **Classification Problem**
    - Definition
    - Loss Functions
    - Learners
2. **Representation Learning**
    - What should be x ?
    - Unsupervised Representation Learning
        - Latent Semantic Analysis
        - GloVE
3. **Demo:** Text Classification with GloVe Embeddings

Classification Problem

Let there be **Data:**

$$(x, y) \sim P_{\mathscr{D}}(X, Y)$$ (unknown)

$y \in \{0,1\}$

$y \in \{0,1,\ldots,k\}$

Deterministic Labeling Function

$$y = f_{gold}(x)$$ (assumption)

Training data

$$S_{train} = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\} \sim P^n_{\mathscr{D}}(x, y)$$ (i.i.d)

Learner

$$\text{Learner}(S_{train}) = f_{learned}$$ predictor

NN + SGD, SVMs

Classification Problem

Loss Function

$$L(\hat{y}, y) = |y - \hat{y}|^2 \qquad \text{(choice)}$$

logistic loss, cross-entropy, max-margin

True Error

$$e(f_{learned}) = \mathbb{E}_{(x,y)\sim P_{\mathscr{D}}}\left[L\left(f_{learned}(x), y\right)\right]$$

population risk

Training Error

$$\hat{e}(f_{learned}) = \frac{1}{n}\sum_{i=1}^{n} L\left(f_{learned}(x), y\right)$$

what learner knows

Empirical Risk Minimization

$$\text{Learner}(S_{train}) = \arg\min_{f_{learned}\in F} \hat{e}(f_{learned})$$

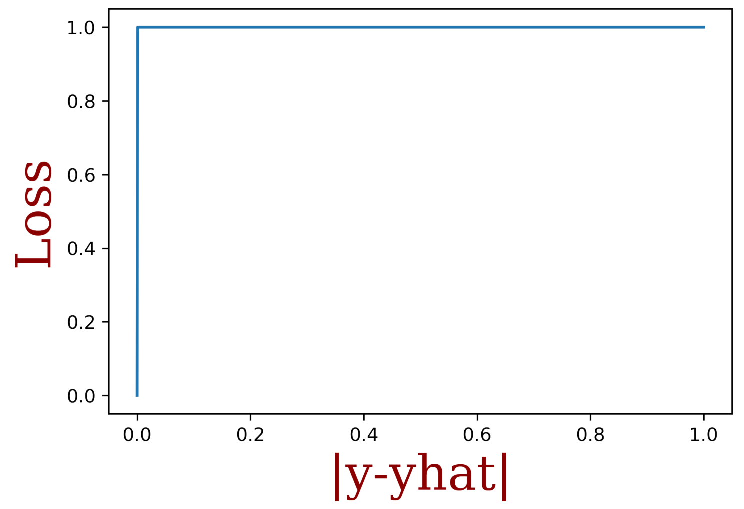
F: parameters of NN

$$f_{learned} = f_{w\in\mathscr{W}}$$

Loss Functions–I: Binary

0–1 Loss

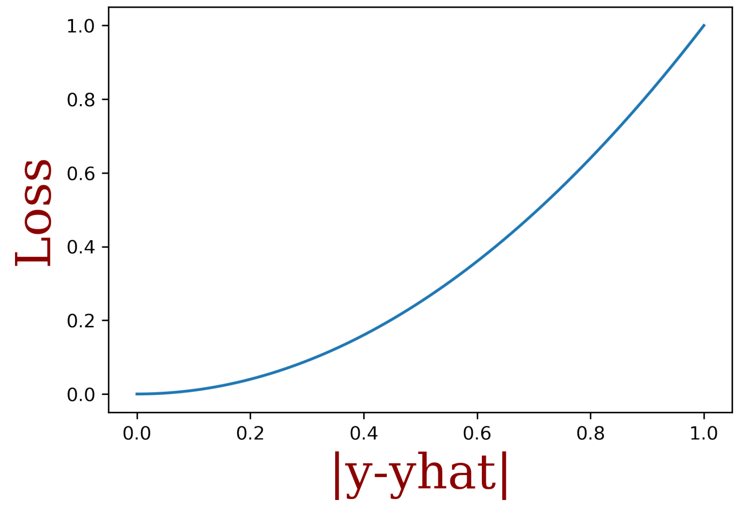
$$L(\hat{y}, y) = 1_{y \neq \hat{y}}$$



$$\hat{y} \in [0,1]$$

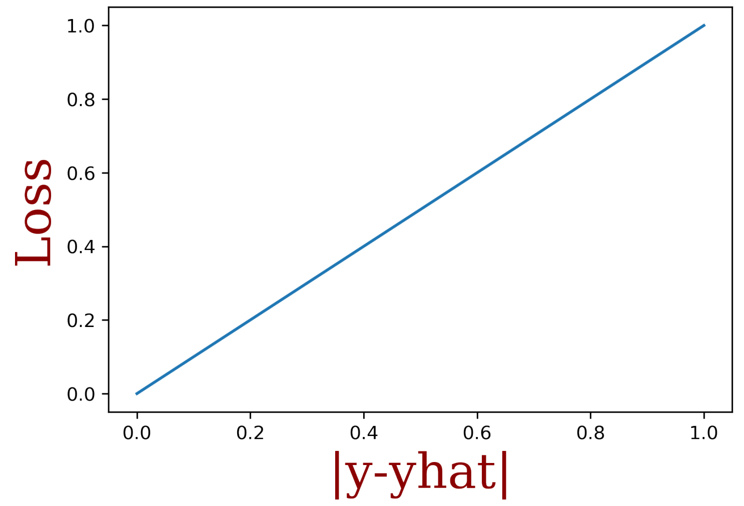
Square Loss

$$L(\hat{y}, y) = |y - \hat{y}|^2$$



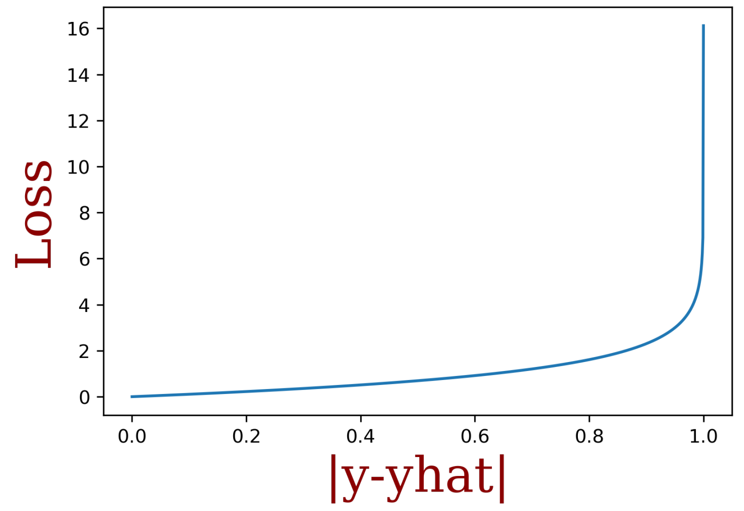
Absolute Loss

$$L(\hat{y}, y) = |y - \hat{y}|$$



Logistic Loss / Binary Cross Entropy Loss

$$L(\hat{y}, y) = y\log(\hat{y}) + (1 - y)\log(1 - \hat{y})$$



Loss Functions-II: Multi-Class

$$f'_{learned}(x) = s \in \mathbb{R}^K$$

$$K = 5 \qquad \overset{s}{\begin{bmatrix} 1.3 \\ 5.1 \\ 2.2 \\ 0.7 \\ 1.1 \end{bmatrix}} \longrightarrow \textbf{softmax}(s) = \frac{\exp(s_y)}{\sum_i \exp(s_i)} \longrightarrow \overset{\hat{p}(y \mid x)}{\begin{bmatrix} 0.02 \\ 0.90 \\ 0.05 \\ 0.01 \\ 0.02 \end{bmatrix}}$$

$$0.90 \rightarrow \hat{p}(y = 1 \mid x)$$

$$s \ \xrightarrow{\arg\max}\ f_{learned}(x) = \hat{y} = 1$$

Loss Functions-II: Multi-Class

$$f'_{learned}(x) = s \in \mathbb{R}^K$$

Cross Entropy Loss / Negative Log Likelihood

$$L(y, s) = -\,log\left(\frac{\exp(s_y)}{\sum_i \exp(s_i)}\right) = \log \textbf{softmax}(s)_y$$

Max-Margin Loss

$$L(y, s) = \max(0, \max(s_{-y}) - s_y + c)$$

Zero-One

$$L(y, s) = 1_{y \neq \arg\max_i s}$$

Learners

<u>Parametric</u>

Neural Networks + SGD

Today's demo

Logistic Regression

used in HW1

SVM

Random forests

<u>Non-Parametric</u>

K-NN

Gaussian Process Classifiers

Representation Learning

input = “this is a sturdy coffee machine , but …”

$$x = ? \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix} \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix} \quad \cdots \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ \vdots \end{bmatrix}$$

think about making learner’s job easy!

– is it easy for classifier to exploit relationships btw. the words in this representation?

– is there a way to learn better input representations?

Unsupervised Representation Learning

If we have access to *unlabeled collection* of inputs, can we learn better representations?

There might be a smaller dense vector space that explain data better than sparse representations?

Latent Semantic Analysis

Term-Document Matrix

is this a good representation?

$$W_{td} = \begin{array}{r} \\ cat \\ dog \\ the \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \left[\begin{array}{ccccccc} \boxed{1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0} \\ 0 & 2 & 0 & 1 & 1 & 1 & 0 \\ 20 & 13 & 18 & 22 & 15 & 4 & 20 \\ & & & & & & \end{array} \right] \end{array}$$

Latent Semantic Analysis

SVD

$$W_{td} = U\Sigma V^T$$

Compact SVD

$$U \in \mathbb{R}^{|w|\times|w|}, \Sigma \in \mathbb{R}^{|w|\times|w|}, V \in \mathbb{R}^{|d|x|w|}$$

assuming $|w| < |d|$

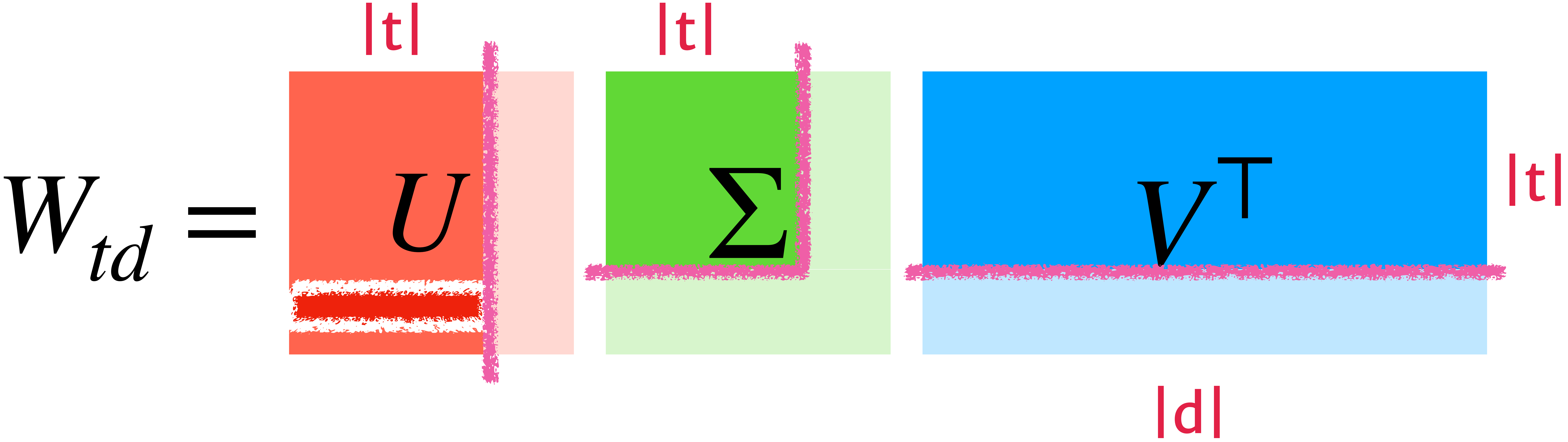
Note that $i < j \implies \sigma_i \geq \sigma_j$

$$= \sum_{i=1}^{|w|} \sigma_i u_i v_i^T \approx \sum_{i=1}^{t} \sigma_i u_i v_i^T$$

(Truncated SVD)

Latent Semantic Analysis

<u>Truncated SVD</u>



Latent Semantic Analysis

Term–Document Matrix

$$
W_{td} = \begin{array}{r c}
 & \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\
\begin{array}{r} cat \\ dog \\ the \\ \\ \end{array} & \left[\begin{array}{ccccccc} 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 & 1 & 1 & 0 \\ 20 & 13 & 18 & 22 & 15 & 4 & 20 \\ & & & & & & \end{array} \right]
\end{array}
$$

Latent Semantic Analysis

TF-IDF

$$TF.IDF = \#(w,d) \times \frac{\#(\text{documents})}{\#(\text{documents has } w)}$$

$$W_{td} = \begin{array}{c} \\ cat \\ dog \\ the \end{array} \begin{array}{c} \begin{array}{ccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 \end{array} \\ \begin{bmatrix} .8 & .8 & 0 & .8 & 0 & .8 & 0 \\ 0 & 2.4 & 0 & 1.1 & 1.1 & 1.1 & 0 \\ \boxed{.02} & \boxed{.01} & \boxed{.02} & \boxed{.02} & \boxed{.01} & \boxed{0} & \boxed{.02} \end{bmatrix} \end{array}$$

GloVe (Global Vectors)

Word2Vec encodes **local** statistics through neighborhood word **prediction**

LSA uses **global** information about word occurence statistics through T-D

GloVe (Global Vectors) captures global information through **word co-occurance matrix**

Remember word co-occurance matrix

$$W_{tt} = \begin{array}{r} \\ cat \\ dog \\ the \end{array} \begin{array}{c} \begin{array}{ccc} cat & dog & the \end{array} \\ \begin{bmatrix} 10 & 8 & 103 \\ 8 & 20 & 97 \\ 103 & 97 & 995 \end{bmatrix} \end{array}$$

GloVe (Global Vectors)

$$P_{k|i} = \frac{X_{ik}}{X_i}$$

Normalize the rows of the word co-occurance matrix

$$P_{k|j} = \frac{X_{jk}}{X_j}$$

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

We should be able to read of this quantity out of word vectors

$$F\left(w_i, w_j, \tilde{w}_k\right) = \frac{P_{k|i}}{P_{k|j}}$$

GloVe

$F\left(w_i, w_j, \tilde{w}_k\right)$ should be a simple function so that we can read from the surface

the information present in $\frac{P_{k|i}}{P_{k|j}}$ is related with semantic difference between word i, and j $\longrightarrow F\left(w_i - w_j, \tilde{w}_k\right) \longrightarrow F\left((w_i - w_j)^T \tilde{w}_k\right)$

if i=j it should 1, if i <=> j it should be 1/F $\longrightarrow F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{F\left(w_i^T \tilde{w}_k\right)}{F\left(w_j^T \tilde{w}_k\right)} = \frac{P_{k|i}}{P_{k|j}}$

$\Longrightarrow F(w_i, \tilde{w}_k) = P_{k|i} = \frac{X_{ik}}{X_i}$ **and** $F = \exp$ **is a solution**

GloVe

if i=j it should 1, if i <=> j it should be 1/F $\rightarrow$ $F\left((w_i - w_j)^T \tilde{w}_k\right) = \dfrac{F\left(w_i^T \tilde{w}_k\right)}{F\left(w_j^T \tilde{w}_k\right)} = \dfrac{P_{k|i}}{P_{k|j}}$

$\Longrightarrow \; F(w_i, \tilde{w}_k) = P_{k|i} = \dfrac{X_{ik}}{X_i}$ **and** $F = \exp$ **is a solution**

$$\Longrightarrow \; w_i^T \tilde{w}_k = \log\left(P_{k|i}\right) = \log\left(X_{ik}\right) - \log\left(X_i\right)$$

$$\Longrightarrow \; w_i^T \tilde{w}_k + \log\left(X_i\right) = \log\left(X_{ik}\right)$$

$$\rightarrow w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log\left(X_{ik}\right)$$

GloVe (Global Vectors)

$$\rightarrow w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log\left(X_{ik}\right)$$

We will globally satisfy this objective!

$$L? = \sum_{i,j=1}^{V} \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2$$

$$L = \sum_{i,j=1}^{V} f\left(X_{ij}\right) \left(\boxed{w_i^T \tilde{w}_j} + b_i + \tilde{b}_j - \boxed{\log X_{ij}} \right)^2$$

(final objective)

Demo!