Introduction to Statistical Machine Learning
CSC/DSCC 265/465

Lecture 7: Supervised Learning – Part IV

Cantay Caliskan

Notes and updates

Notes and updates

- The deadline for the 3rd Problem Set is **Friday, February 11, 11:59 PM!**

Plan for today

- ***Model: Logistic Regression***
- ***Bias and Variance***
- ***Cross-Validation***

Plan for today

- ***Model: Logistic Regression***
- ***Bias and Variance***
- ***Cross-Validation***

Technical Background of Logistic Regression

Logistic Regression: Hypothesis

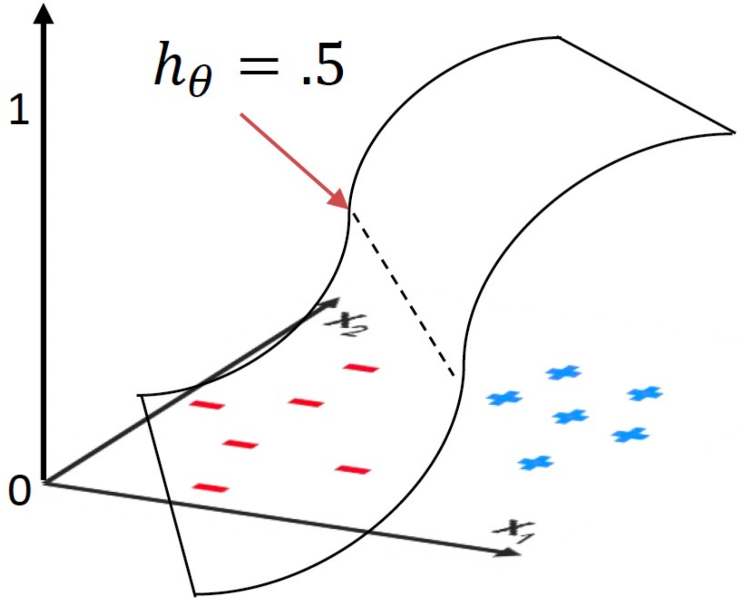
- <u>Hypothesis</u>:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

- Predict **y = 1** if $h_\theta(x) \geq 0.5$
- Predict **y = 0** if $h_\theta(x) < 0.5$

- <u>Cost function</u>: **Cross-entropy**

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$
$$= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$



Separating hyperplane is n-dimensional (where n=number of features)

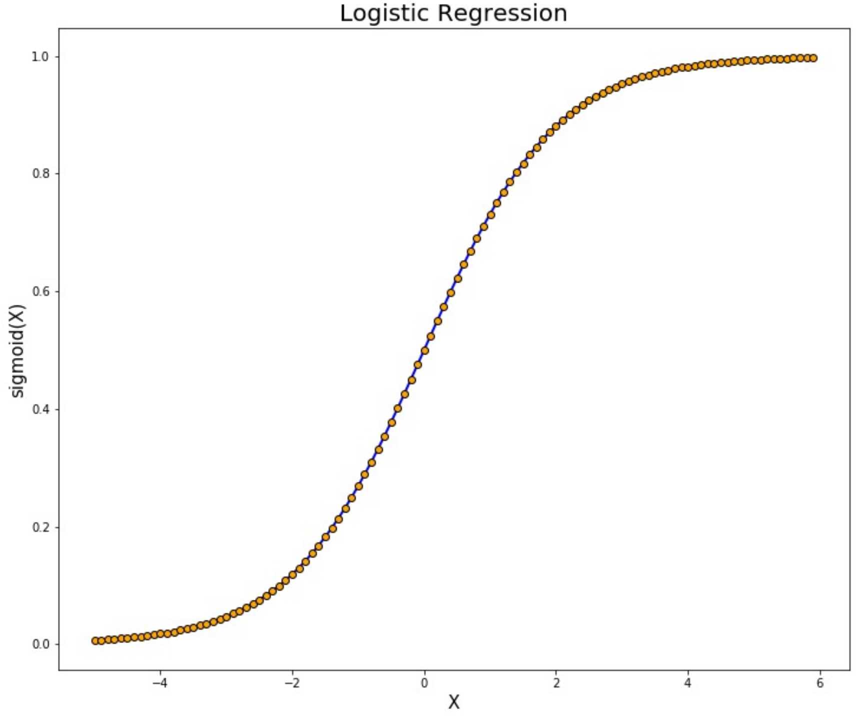
<u>Goal</u>: Minimize $J(\theta)$

<u>Question</u>: What does cross-entropy measure?

UNIVERSITY of ROCHESTER

Logistic Regression

- $0 \leq h_\theta(x) \leq 1$
- Idea: Using a ***sigmoid*** function, map the input vector to (0,1)



$$g(z) = \frac{1}{1+e^{-z}}$$

Sigmoid function

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

Sigmoidal hypothesis function (generalized for logistic regression)

$$h_\theta(x) = p(y=1|x)$$

Hypothesis: Probability of 'being member of class 1' given input

Sigmoid function

- ***First***: We need to create a linear combination ***z***:

$$z = \left(\sum_{i=1}^{n} w_i x_i\right) + b$$

$$z = (\boldsymbol{w} * \boldsymbol{x}) + b$$

<u>Question</u>: What can the range for *z* be?

<u>Answer</u>: $-\infty\ to + \infty$

- *Second*: We will pass ***z*** through the Sigmoid function ***S***:

$$S(z) = \frac{1}{1 + e^{-z}}$$

Sigmoid function

- ***Third***: Apply the Sigmoid function to the sum of weighted features:

$$P(y=1) = S(\boldsymbol{w}.\boldsymbol{x} + b)$$

$$P(y=1) = \frac{1}{1 + e^{(-(\boldsymbol{w}*\boldsymbol{x}+b))}}$$

$$P(y=0) = 1 - S(\boldsymbol{w}.\boldsymbol{x} + b)$$

$$P(y=0) = 1 - \frac{1}{1 + e^{(-(\boldsymbol{w}*\boldsymbol{x}+b))}}$$

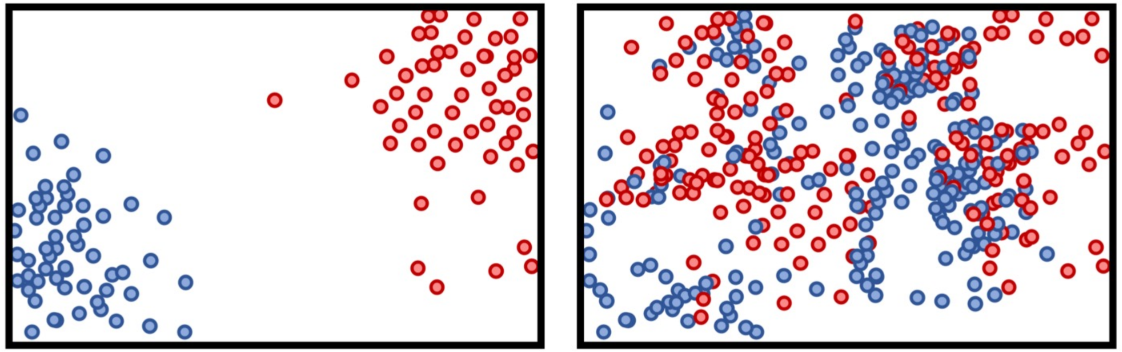
Question: What is this value called?

Note: Another property: $1 - S(z) = S(-z)$

Question: How can we express P(y=0)?

Entropy and Cross-Entropy

- ***Entropy***: A measure that quantifies the 'state of disorder'
  - In ML: A measure that quantifies the **purity** of a classification result

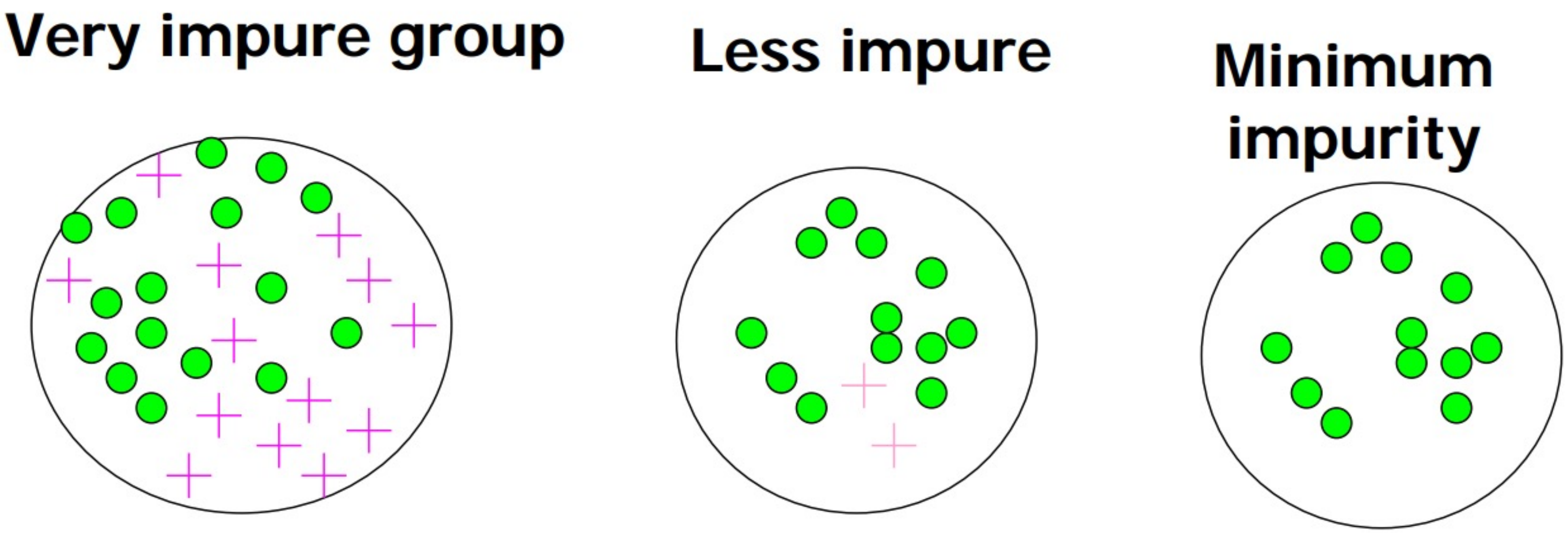


Low Entropy High Entropy

- ***Cross-entropy***: A measure of the relative entropy between two probability distributions over the same set of events
  - In Logistic Regression: Labels are associated with probabilities
    - Example: For two-class logistic regression (***p*** and ***q***)

Entropy

- A mathematical formula that measures the level if ***impurity*** in a group of examples



- Formula for <u>entropy</u>: $-\sum_{i=1}^{K} p_k * log_2\,(p_k)$ where p_k denotes the proportion of instances belonging to class ***k*** (*K=1, …, k*).

Example: Entropy

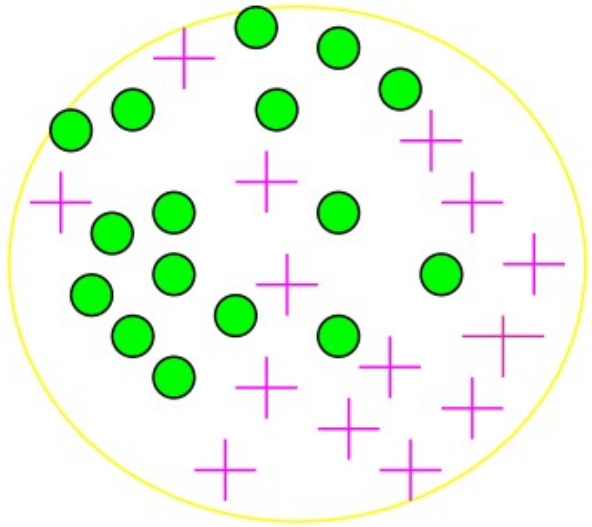
- *Entropy*: $-\sum_{i=1}^{K} p_k * log_2$

- Let's say we have a cloud of points

- Let's compute the proportion of class k in the set:

- 16/30 are **green circles**; 14/30 are **pink crosses**
- $log_2\left(\frac{16}{30}\right) = -.9\ and\ log_2\left(\frac{14}{30}\right) = -1.1$
- Entropy: (-16/30)*(-.9) - (14/30)*(-1.1) = 0.99

- The ***higher*** the entropy, the ***less pure*** the information content
  - And the ***better*** our set for learning!



<u>Properties:</u>

Entropy is **maximized** when elements are heterogenous (impure)

Entropy is **minimized** when elements are homogenous (pure)

Best Parameters

- Clarification: In the logistic regression model, each class is associated with a (linear combination) of ***log-odds*** (logarithm of the odds) value
  - Question: Why?
  - Answer: Goodness of fit is measured by the likelihood function and best parameters are obtained when log-likelihood is *maximized*
- Specifically, (For a ***two-class*** logistic function):

$$L = \prod_{i:y_i=1} p_i \prod_{i:y_i=0} (1 - p_i)$$

We need to maximize the ***likelihood*** function.

$$\ell = \sum_{i:y_i=1} \ln(p_i) + \sum_{i:y_i=0} \ln(1-p_i) = \sum_i [y_i \ln(p_i) + (1-y_i)\ln(1-p_i)]$$

Which corresponds to maximizing ***log-likelihood***.

$$0 = \frac{\partial \ell}{\partial \beta_0} = \sum_i (y_i - p_i)$$

$$0 = \frac{\partial \ell}{\partial \beta_1} = \sum_i (y_i - p_i) x_i$$

Best parameters for our coefficients are found.

Question: How do we find the best parameters?

UNIVERSITY of ROCHESTER

Converting Log-Odds to Probability

- Question: Why do we convert from ***log odds*** to ***probability***?
- Answer: It is difficult to model a variable when it has a restricted range
  - And, probability has a restricted range. Log-odds doesn't.
- Transformation from log-odds to probability is called ***logit transformation***
  - And, sometimes *logistic regression* is called ***logit regression***

$$logit(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Log-odds (i)

$$\frac{1-p}{p} = \frac{1}{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}$$

Exponentiate by ***e*** and take the multiplicative inverse (ii)

$$\frac{1}{p} = 1 + \frac{1}{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}$$

Add **1** to both sides (iii)

$$\frac{1}{p} = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) + 1}{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}$$

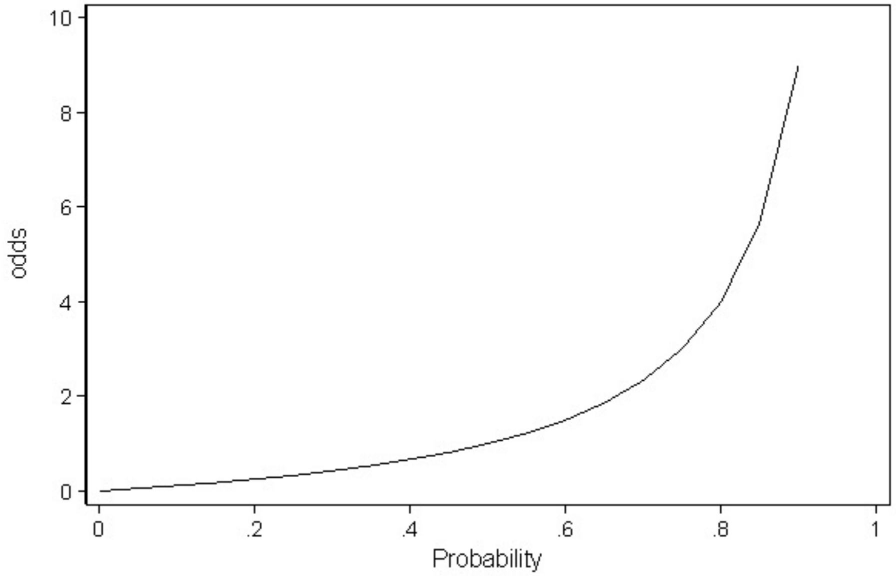
Add a common denominator(iv)

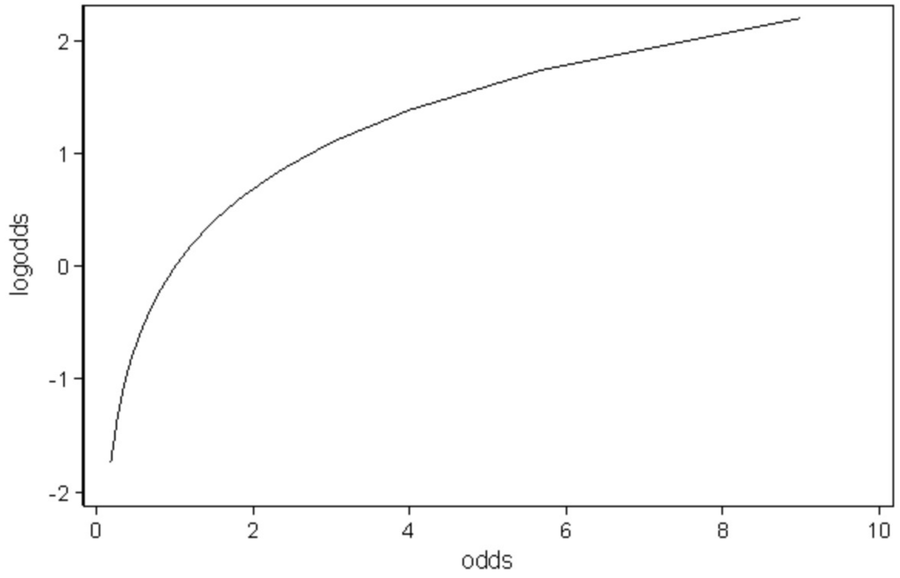
$$p = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}$$

Take the multiplicative inverse again to find the probability

Example: Converting Log-Odds to Probability

- Idea: Let's say probability of success is ***s***, and probability of failure is ***f***
  - And ***s*** = 0.8 and ***f*** = 0.2
  - Odds of success: 0.8/0.2 = 4
  - Log odds of success log(0.8/0.2) = log(4)
  - Probabilistically speaking: 62.07%
  - Note: Probability ranges from 0 to 1, odds range from 0 to ∞





Summary: Logistic Function

- What are the different components?

  1) ***Input***: A feature representation of the input: $[x_1, x_2, \ldots, x_n]$

  2) *Function*: A classification function that computes **ŷ** via ***P(y|x)***

  Logistic function uses ***sigmoid*** or ***softmax*** function for classification

  3) ***Cost***: A cost function for learning that helps to minimize error on training set

  Logistic function uses ***cross-entropy*** cost function

  4) ***Optimization***: An algorithm for optimizing the cost function

  One option: ***Stochastic Gradient Descent*** algorithm

  5) ***Result***: Best parameters are log-odds with with respect to a reference category

  Reference category is usually the 'most usual' / average / expected one

UNIVERSITY of ROCHESTER

Let's go through an example and a practical application…

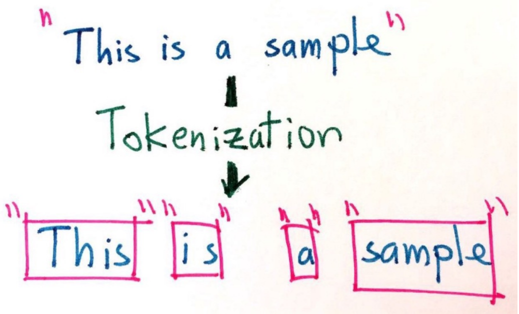
Example: Sentiment Classification

- Let’s work out a simple <u>sentiment classification</u>:
  - Suppose we are doing a binary sentiment classification on a movie review text, and we would like assign the sentiment class + or – to a review **document**.

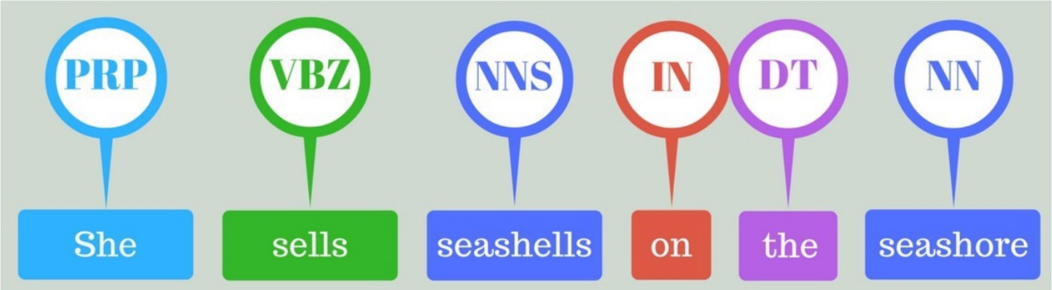
Positive	**Negative**
GREAT movie and the family will love it!! If kids are bored one day just pop the tape in and you'll be so glad you did!!! ~~~Rube i luv raven-s!	The script for this movie was probably found in a hair-ball recently coughed up by a really old dog. Mostly an amateur film with lame FX. For you Zeta-Jones fanatics: she has the credibility of one Mr. Binks.
Did Sandra (yes, she must have) know we would still be here for her some nine years later? See it if you haven't, again if you have; see her live while you can.	I would love to have that two hours of my life back. It seemed to be several clips from Steve's Animal Planet series that was spliced into a loosely constructed script. Don't Go, If you must see it, wait for the video ...
Verry classic plot but a verry fun horror movie for home movie party Really gore in the second part This movie proves that you can make something fun with a small budget. I hope that the director will make another one	This is without a doubt the worst movie I have ever seen. It is not funny. It is not interesting and should not have been made.

Example: Sentiment Classification

1) Consider ***cleaning the data***: Remove HTML links, stopwords, emojis, non-alphabetic characters, programming code, words that are too short
2) Consider ***tokenization*** and ***part-of-speech-tagging*** and ***stemming*** and ***lemmatization***:
   - ***Tokenization***: Break down a sentence to understand the functions of individual words



   - ***POS tagging***: Labeling the words in a text according to their Word types (noun, adjective, adverb, verb etc.)



Oh, we need another **supervised model** here!

Example: Sentiment Classification

3) Consider ***tokenization*** and ***part-of-speech-tagging*** and ***stemming*** and ***lemmatization*** :

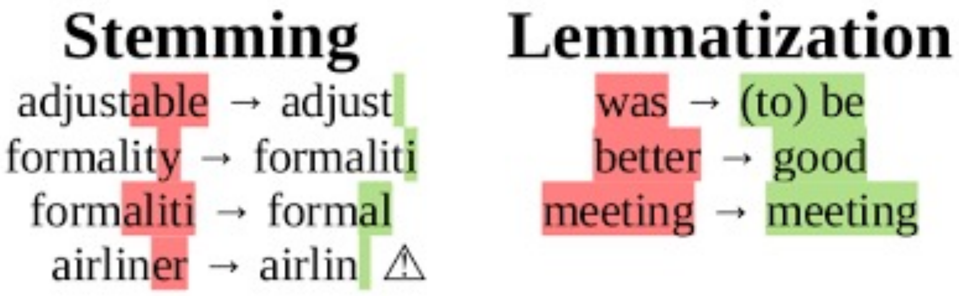
- ***Stemming***: The process of finding the root of the words

changing, changed, change → *stemming* → chang, chang, chang

Question: Is there a best stemming method?

studying, studies, study → *stemming* → studi, studi, studi

- ***Lemmatization***: The process of finding the form of the related word in the dictionary (not equal to stemming)
  - Use lexical knowledge instead of base of words



Example: Sentiment Classification

- How can ve ***vectorize*** the word representations?
  1) ***Bag of Words***: Extract all the unique words in the corpus and then count how many times a particular word appears

```
It was the best of times,

it was the worst of times,

it was the age of wisdom,

it was the age of foolishness.

Unique words:
"it", "was", "the", "best", "of", "times", "worst", "age", "wisdom"
"foolishness"
```

Example: Sentiment Classification

- How can ve ***vectorize*** the word representations?
  1) ***Bag of Words***: Extract all the unique words in the corpus and then count how many times a particular word appears

```
"it was the worst of times" = [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]
```

```
"it was the age of wisdom" = [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]
```

```
"it was the age of foolishness" = [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]
```

Question: What is the disadvantage here?

Example: Sentiment Classification

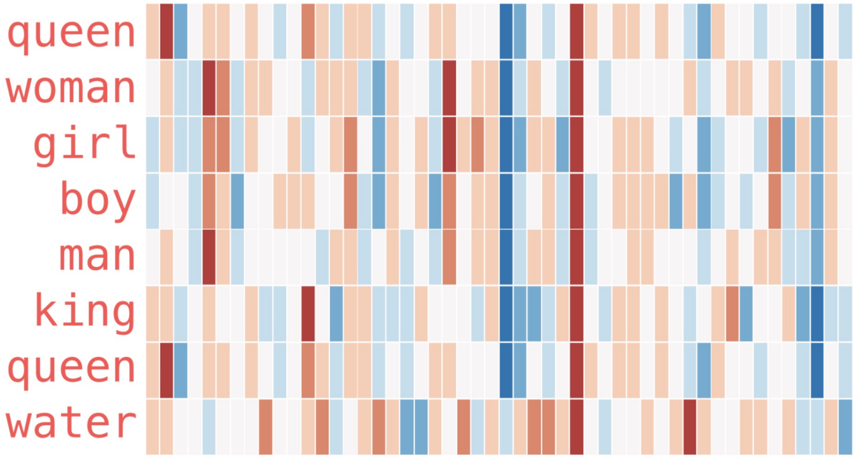
- How can ve ***vectorize*** the word representations?

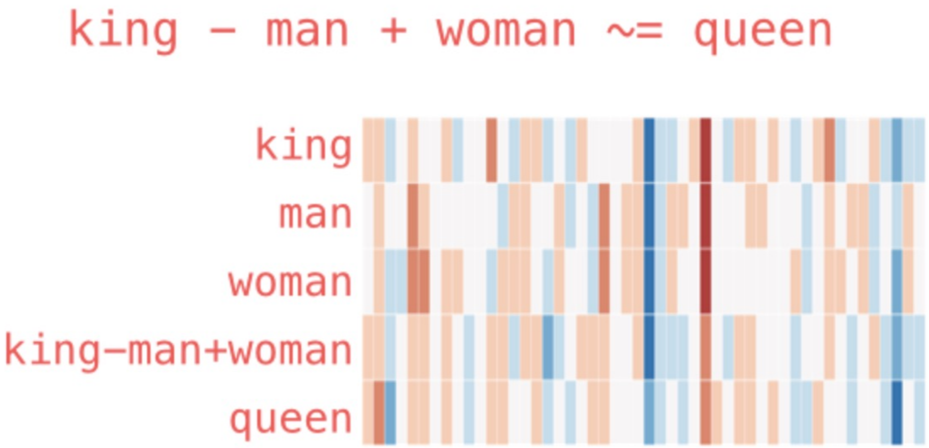
  2) ***Word2Vec***: A 2013 algorithm invented by a Google employee

  - Idea: Have words occupy spatially close vector spaces.

    Create a vector space for documents by using:

    - context(s) for the predicting word (CBOW)
    - word(s) for predicting the context (skip-gram)





UNIVERSITY of ROCHESTER

Example: Sentiment Classification

- How can ve ***vectorize*** the word representations?

  3) ***TF-IDF***: Term-frequency-inverse-document-frequency

     - Idea: Calculate the measure of how important a Word is in the document to calculate the importance of a document

$$Tf = \frac{number\ of\ times\ term\ appears\ in\ document}{total\ number\ of\ words\ in\ document}$$

$$Idf = \ln\left(\frac{total\ number\ of\ documents}{number\ of\ documents\ with\ term\ in\ them}\right)$$

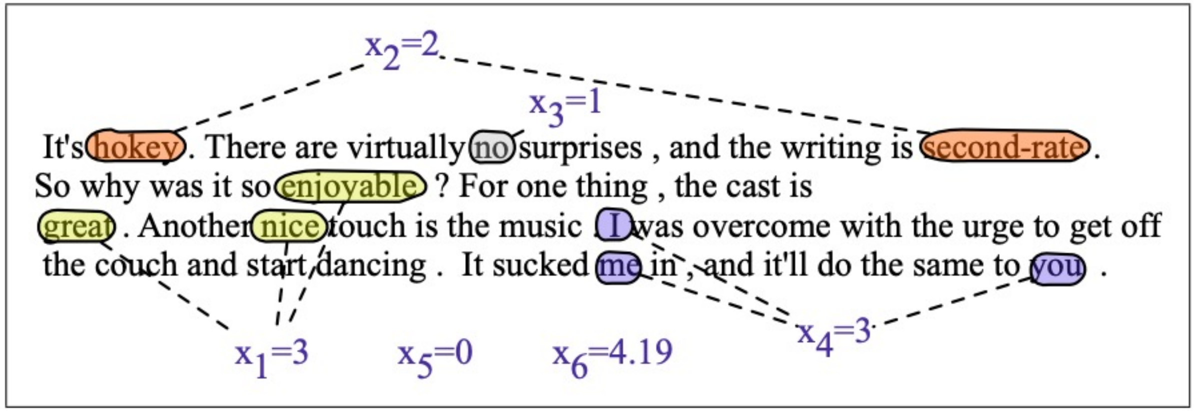
$$TfIdf = Tf \times Idf$$

Example: Sentiment Classification

- How do we compute the ***classes***?

  By using ***linear combinations*** and the ***sigmoid function***:

Var	Definition	Value in Fig. 5.2
x_1	count(positive lexicon words $\in$ doc)	3
x_2	count(negative lexicon words $\in$ doc)	2
x_3	$\begin{cases} 1 & \text{if "no"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	1
x_4	count(1st and 2nd pronouns $\in$ doc)	3
x_5	$\begin{cases} 1 & \text{if "!"} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	0
x_6	log(word count of doc)	$\ln(66) = 4.19$



Let's say ***w_1 = 2,5*** and ***w_2 = -5:***

$$
\begin{aligned}
p(+|x) = P(y=1|x) &= \sigma(\mathbf{w}\cdot\mathbf{x}+b) \\
&= \sigma([2.5,-5.0,-1.2,0.5,2.0,0.7]\cdot[3,2,1,3,0,4.19]+0.1) \\
&= \sigma(.833) \\
&= 0.70 \qquad (5.7) \\
p(-|x) = P(y=0|x) &= 1-\sigma(\mathbf{w}\cdot\mathbf{x}+b) \\
&= 0.30
\end{aligned}
$$

Question: What is the vectorization approach here?

Source: Jurafsky

UNIVERSITY of ROCHESTER