Text Classification II

Sandeep Soni

09/27/2023

Class logistics

- PS1 due on 10/2
  - No collaboration allowed
- Project proposal due 10/11
- Office hours today between 4-5pm

PS1 Issues

PS1 Issues

- In 1.2, for calculating vocabulary richness using the characteristic measure, the equation is $\frac{1}{N^2}\left(\left(\sum_{i=1}^{N} i^2 V_i\right) - N\right)$

PS1 Issues

- In 1.2, for calculating vocabulary richness using the characteristic measure, the equation is $\frac{1}{N^2}\left(\left(\sum_{i=1}^{N} i^2 V_i\right) - N\right)$
- This value should not be negative! (See piazza)

PS1 Issues

- In 1.2, for calculating vocabulary richness using the characteristic measure, the equation is $\frac{1}{N^2}\left(\left(\sum_{i=1}^{N} i^2 V_i\right) - N\right)$
- This value should not be negative! (See piazza)
- I also asked to calculate richness on simple examples
  - $V = \{1:1, 2:1, 3:1, 4:1, 5:1\}$, N=15 (should evaluate to 0.178)
  - $V = \{1:2, 2:2, 3:0, 4:0, 5:1\}$, N=11 (should evaluate to 0.198)

PS1 Issues

PS1 Issues

- In 1.3, for calculating vocabulary richness using the entropy, the base of the logarithm is typically 2 but in my sanity check I have used natural logarithm (base e)

PS1 Issues

- In 1.3, for calculating vocabulary richness using the entropy, the base of the logarithm is typically 2 but in my sanity check I have used natural logarithm (base e)
- You are free to use any base in your calculation

Project proposal

- 2 pages
- Include at least 5 references
- Submit one proposal per group
- Please mention team member names in the proposal

- Proposal should answer the following questions
  - What is the research question?
  - How has it been answered before?
  - Why is the question interesting?
  - What data will you use to answer this question?
  - What methods/tools will you use or build?
  - How will you evaluate your success?
  - What is your timeline?

Example Proposal Sketch

Question	Example
What is/are the research question(s)?	Is gender bias increasing in songs?
How has it been answered before?	through qualitative work, text analysis only on English songs, etc
Why is the question interesting?	helps understand if objectification or stereotyping is prevalent, rising, or declining
What data will you use to answer the question(s)?	song lyrics from 3 languages, etc
What methods/tools will you use or build?	use combination of word embeddings and classification techniques, etc
How will you evaluate your success?	compare against baselines, perform statistical tests, etc
What is your timeline?	data collection by Oct, execution by Nov, etc

Question for the day

"How to predict a label for given text?"

Agenda

- Review Naive Bayes
- Logistic regression
- Evaluation techniques
- Applications

Supervised classification

Supervised classification

- Objective
  - There exists h such that $y = h(x)$, where $x \in \mathcal{X}$, $y \in \mathcal{Y}$
  - Find $\hat{h}$ that closely approximates h

Supervised classification

- Objective
  - There exists h such that $y = h(x)$, where $x \in \mathcal{X}$, $y \in \mathcal{Y}$
  - Find $\hat{h}$ that closely approximates h
- How?
  - Assuming $\hat{h}$ has some parametric form, learn parameters from training data of the form $\{x^{(j)}, y^{(j)}\}_{j=1}^{M}$

Supervised classification

Supervised classification

- For one example, the conditional probability of the label given the text is:

Supervised classification

- For one example, the conditional probability of the label given the text is:

- $$P(y^{(j)} \mid x^{(j)}) = \frac{P(y^{(j)}, x^{(j)})}{P(x^{(j)})}$$

Supervised classification

- For one example, the conditional probability of the label given the text is:

- $$P(y^{(j)} \mid x^{(j)}) = \frac{P(y^{(j)}, x^{(j)})}{P(x^{(j)})}$$

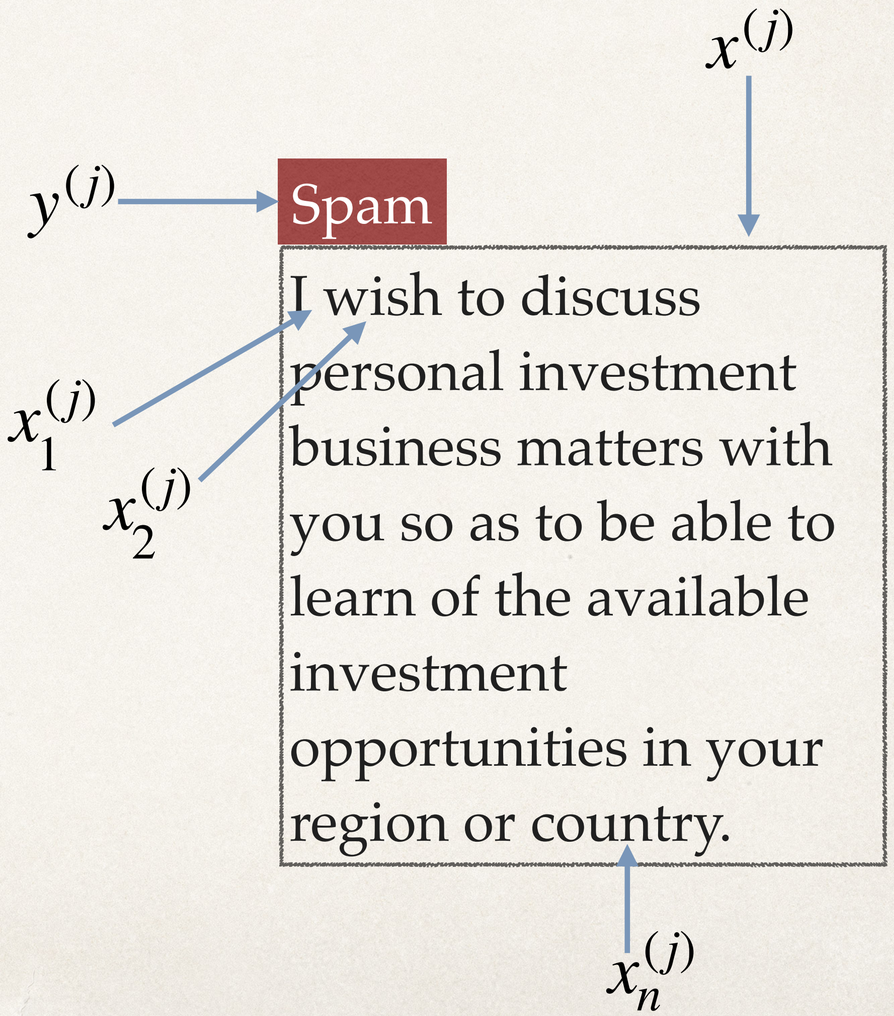
- $$P(y^{(j)} \mid x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}) = \frac{P(y^{(j)}, x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)})}{P(x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)})}$$

Supervised classification

- For one example, the conditional probability of the label given the text is:

- $P(y^{(j)} | x^{(j)}) = \dfrac{P(y^{(j)}, x^{(j)})}{P(x^{(j)})}$

- $P(y^{(j)} | x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}) = \dfrac{P(y^{(j)}, x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)})}{P(x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)})}$



Chain Rule

$$P(y, x_1, x_2, \ldots, x_n) \propto P(y)P(x_1 \mid y)P(x_2 \mid x_1, y)P(x_3 \mid x_2, x_1, y)\ldots P(x_n \mid x_{n-1}, \ldots, x_1, y)$$

Chain Rule

$$P(y, x_1, x_2, \ldots, x_n) \propto P(y)P(x_1 \mid y)P(x_2 \mid x_1, y)P(x_3 \mid x_2, x_1, y)\ldots P(x_n \mid x_{n-1}, \ldots, x_1, y)$$

- Estimation is intractable because too many parameters
  - For binary labels, $P(y)$ is 1 parameter
  - For binary features, $P(x_1 \mid y)$ needs 2 parameters
  - $P(x_2 \mid x_1, y)$ needs 4 parameters
  - $P(x_3 \mid x_2, x_1, y)$ needs 8 parameters

Chain Rule

$$P(y, x_1, x_2, \ldots, x_n) \propto P(y)P(x_1 \mid y)P(x_2 \mid x_1, y)P(x_3 \mid x_2, x_1, y)\ldots P(x_n \mid x_{n-1}, \ldots, x_1, y)$$

P(y=spam)

P(“I” | y=spam)
P(“I” | y=~spam)

P(“wish” | ”I”, spam)
P(“wish” | ~”I”,spam)
P(“wish” | ”I”, ~spam)
P(“wish” | ~”I”,~spam)

Chain Rule

$$P(y, x_1, x_2, \ldots, x_n) \propto P(y)P(x_1 \mid y)P(x_2 \mid x_1, y)P(x_3 \mid x_2, x_1, y)\ldots P(x_n \mid x_{n-1}, \ldots, x_1, y)$$

- Estimation is intractable because too many parameters
  - For binary labels, $P(y)$ is 1 parameter
  - For binary features, $P(x_1 \mid y)$ needs 2 parameters
  - $P(x_2 \mid x_1, y)$ needs 4 parameters
  - $P(x_3 \mid x_2, x_1, y)$ needs 8 parameters …

P(y=spam)

P(“I” | y=spam)
P(“I” | y=~spam)

P(“wish” | ”I”, spam)
P(“wish” | ~”I”,spam)
P(“wish” | ”I”, ~spam)
P(“wish” | ~”I”,~spam)

Naive Bayes

$$P(y \mid x_1, x_2, \ldots, x_n) \propto P(y)P(x_1 \mid y)P(x_2 \mid y)P(x_3 \mid y)\ldots P(x_n \mid y)$$

P(“I” | y=spam)

P(“I” | y=~spam)

P(y=spam)

P(“wish” | y=spam)

P(“wish” | y=~spam)

P(“country” | y=spam)

P(“country” | y=~spam)

Naive Bayes

$$P(y \mid x_1, x_2, \ldots, x_n) \propto P(y)P(x_1 \mid y)P(x_2 \mid y)P(x_3 \mid y)\ldots P(x_n \mid y)$$

- If we make the conditional independence assumption of Naive Bayes, estimation becomes tractable

P(y=spam)

P(“I” | y=spam)
P(“I” | y=~spam)

P(“wish” | y=spam)
P(“wish” | y=~spam)

P(“country” | y=spam)
P(“country” | y=~spam)

Naive Bayes

Naive Bayes

- Let’s rewrite the posterior probability under naive Bayes

Naive Bayes

- Let's rewrite the posterior probability under naive Bayes

$$P(y|x) \propto P(y)P(x_1|y)P(x_2|y)P(x_3|y)\ldots P(x_n|y)$$

Naive Bayes

- Let's rewrite the posterior probability under naive Bayes

$$P(y|x) \propto P(y)P(x_1|y)P(x_2|y)P(x_3|y)\ldots P(x_n|y)$$

$$\propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

Naive Bayes

- Let's rewrite the posterior probability under naive Bayes

$$P(y|x) \propto P(y)P(x_1|y)P(x_2|y)P(x_3|y)\ldots P(x_n|y)$$
$$\propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

- If we take a log of both sides, we get

Naive Bayes

- Let's rewrite the posterior probability under naive Bayes

$$P(y|x) \propto P(y)P(x_1|y)P(x_2|y)P(x_3|y)\ldots P(x_n|y)$$
$$\propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

- If we take a log of both sides, we get

$$\log P(y|x) \propto \log P(y) + \sum_{i=1}^{N} \log P(x_i|y)$$

$$\log P(y \mid x) \propto \log P(y) + \sum_{i=1}^{N} \log P(x_i \mid y)$$

$$\log P(y \mid x) \propto \log P(y) + \sum_{i=1}^{N} \log P(x_i \mid y)$$

$$\log P(y \mid x) \propto 1 \cdot \log P(y) + \sum_{i=1}^{N} 1 \cdot \log P(x_i \mid y)$$

$$\log P(y \mid x) \propto \log P(y) + \sum_{i=1}^{N} \log P(x_i \mid y)$$

$$\log P(y \mid x) \propto 1 \cdot \log P(y) + \sum_{i=1}^{N} 1 \cdot \log P(x_i \mid y)$$

$$\log P(y \mid x) \propto f_0 \cdot w_0 + \sum_{i=1}^{N} f_i \cdot w_i$$

$$\log P(y \mid x) \propto \log P(y) + \sum_{i=1}^{N} \log P(x_i \mid y)$$

$$\log P(y \mid x) \propto 1 \cdot \log P(y) + \sum_{i=1}^{N} 1 \cdot \log P(x_i \mid y)$$

$$\log P(y \mid x) \propto f_0 \cdot w_0 + \sum_{i=1}^{N} f_i \cdot w_i$$

$$\log P(y \mid x) \propto \sum_{i=0}^{N} f_i \cdot w_i = w^T f$$

$$\log P(y \mid x) \propto \log P(y) + \sum_{i=1}^{N} \log P(x_i \mid y)$$

$$\log P(y \mid x) \propto 1 \cdot \log P(y) + \sum_{i=1}^{N} 1 \cdot \log P(x_i \mid y)$$

$$\log P(y \mid x) \propto f_0 \cdot w_0 + \sum_{i=1}^{N} f_i \cdot w_i$$

Weights

$$\log P(y \mid x) \propto \sum_{i=0}^{N} f_i \cdot w_i = w^T f$$

Features

$$\log P(y \mid x) \propto \log P(y) + \sum_{i=1}^{N} \log P(x_i \mid y)$$

$$\log P(y \mid x) \propto 1 \cdot \log P(y) + \sum_{i=1}^{N} 1 \cdot \log P(x_i \mid y)$$

$$\log P(y \mid x) \propto f_0 \cdot w_0 + \sum_{i=1}^{N} f_i \cdot w_i$$

Naive Bayes is a linear model

Weights

$$\log P(y \mid x) \propto \sum_{i=0}^{N} f_i \cdot w_i = w^T f$$

Features

Logistic Regression

Logistic Regression

In logistic regression, we directly model the conditional probability of the label given the text

Logistic Regression

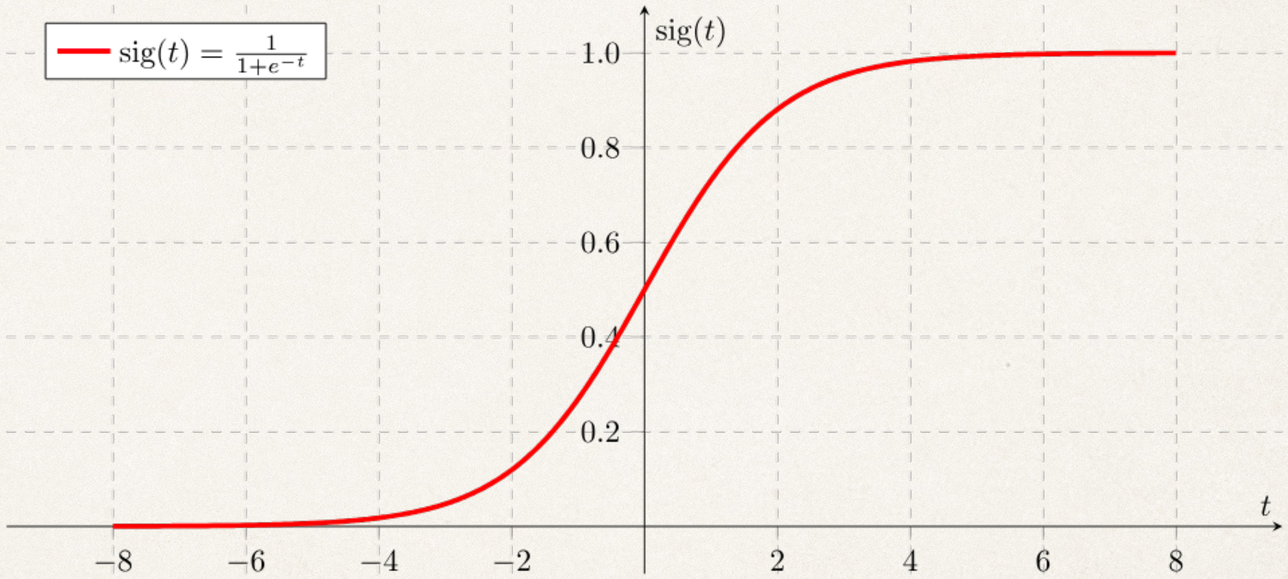
In logistic regression, we directly model the conditional probability of the label given the text

$$P(y \mid x) = \mathrm{sig}(w^T f(x))$$

Logistic Regression

In logistic regression, we directly model the conditional probability of the label given the text

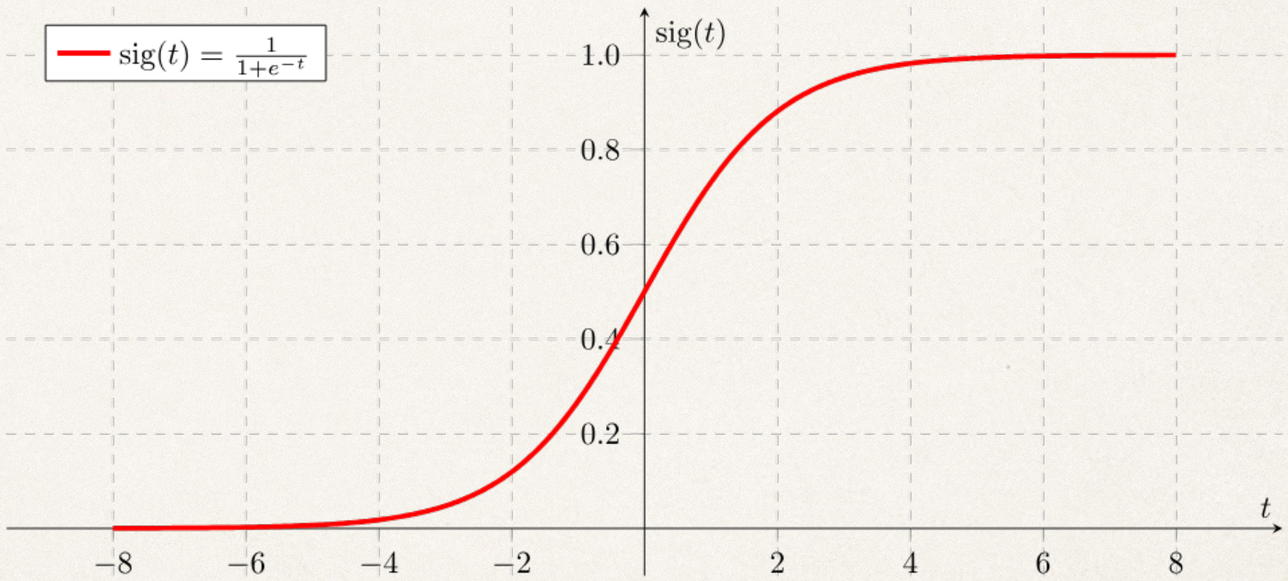
$$P(y \mid x) = \mathrm{sig}(w^T f(x))$$



Logistic Regression

In logistic regression, we directly model the conditional probability of the label given the text

$$P(y \mid x) = \text{sig}(w^T f(x))$$



This is highly flexible because we can encode any type of features that we think could be useful

Learning

Learning

- The objective is to learn the vector of weights w

Learning

- The objective is to learn the vector of weights w
- Unlike Naive Bayes, there is no closed form solution

Learning

- The objective is to learn the vector of weights w
- Unlike Naive Bayes, there is no closed form solution
- Instead we learn the weights by framing the classification task as an optimization problem and estimating the weights using an iterative algorithm applied on the training data

Generalized classification

Generalized classification

- Linear models are not the only models we can use

Generalized classification

- Linear models are not the only models we can use
- We can think of features being transformed by some non-linear function g which could be composed of many functions, so …

Generalized classification

- Linear models are not the only models we can use
- We can think of features being transformed by some non-linear function g which could be composed of many functions, so …
- $P(y \mid x) = \text{sig}(g(f(x)))$

Generalized classification

- Linear models are not the only models we can use
- We can think of features being transformed by some non-linear function g which could be composed of many functions, so …
- $P(y \mid x) = \text{sig}(g(f(x)))$ g could be learned using a neural network

Generalized classification

- Linear models are not the only models we can use
- We can think of features being transformed by some non-linear function g which could be composed of many functions, so …
- $P(y \mid x) = \text{sig}(g(f(x)))$ g could be learned using a neural network
- For multi-class classification, $P(y \mid x) = \text{softmax}(g(f(x)))$

Forecasting

Forecasting

- Instead of converting the scores into probabilities, the same models can be used to predict a continuous value from text
- $y = g(f(x))$
- If g is a linear function, this is linear regression

Feature function

Feature function

- Feature function transforms text into a feature vector
  - e.g. text —> counts vector
  - e.g. text —> counts vector | topics proportion

Regularization

Regularization

- Often you may want to enforce some structure to the weights that you learn (e.g., the weights should not be too high)

Regularization

- Often you may want to enforce some structure to the weights that you learn (e.g., the weights should not be too high)
- One way to enforce such structure is by adding some penalty term to the optimization objective when learning parameters of the model

Regularization

- Often you may want to enforce some structure to the weights that you learn (e.g., the weights should not be too high)
- One way to enforce such structure is by adding some penalty term to the optimization objective when learning parameters of the model
- The value of the penalty is controlled by a hyper parameter which should be tuned on the development set

How do we know if our learned classifier is good?

	1	2	3	4	5	6	7	8	9	10
y	spam	spam	spam	spam	spam	ham	ham	ham	ham	ham
yhat	spam	spam	ham	spam	spam	spam	ham	spam	ham	spam

	1	2	3	4	5	6	7	8	9	10
y	spam	spam	spam	spam	spam	ham	ham	ham	ham	ham
yhat	spam	spam	ham	spam	spam	spam	ham	spam	ham	spam

Confusion Matrix

Confusion Matrix

- On test data, we can tabulate a confusion matrix

Confusion Matrix

- On test data, we can tabulate a confusion matrix

Predicted\Observed	y=spam	y=ham
yhat = spam	4	3
yhat=ham	1	2

Confusion Matrix

Confusion Matrix

- On test data, we can tabulate a confusion matrix

Confusion Matrix

- On test data, we can tabulate a confusion matrix

Predicted\Observed	y=spam	y=ham
yhat = spam	True positives	False positives
yhat=ham	False negatives	True negatives

Confusion Matrix

- On test data, we can tabulate a confusion matrix

Predicted\Observed	y=spam	y=ham
yhat = spam	True positives	False positives
yhat=ham	False negatives	True negatives

- N = True positives + True negatives + False positives + False negatives

Accuracy

Accuracy

- Accuracy

Predicted\Observed	y=spam	y=ham
yhat = spam	True positives	False positives
yhat=ham	False negatives	True negatives

Accuracy

- Accuracy

$$\frac{\text{True positives} + \text{True negatives}}{N}$$

Predicted\Observed	y=spam	y=ham
yhat = spam	True positives	False positives
yhat=ham	False negatives	True negatives

Accuracy

- Accuracy

$$\frac{\text{True positives} + \text{True negatives}}{N}$$

Predicted\Observed	y=spam	y=ham
yhat = spam	True positives	False positives
yhat=ham	False negatives	True negatives

- We want the accuracy of the classifier to be high

Precision

Precision

- Precision

Predicted\Observed	y=spam	y=ham
yhat = spam	True positives	False positives
yhat=ham	False negatives	True negatives

Precision

- Precision

$$\frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

Predicted\Observed	y=spam	y=ham
yhat = spam	True positives	False positives
yhat=ham	False negatives	True negatives

Precision

- Precision

$$\frac{\text{True positives}}{\text{True positives + False positives}}$$

Predicted\Observed	y=spam	y=ham
yhat = spam	True positives	False positives
yhat=ham	False negatives	True negatives

- We want number of false positives to be low and precision to be high

Recall

Recall

- Recall

Predicted\Observed	y=spam	y=ham
yhat = spam	True positives	False positives
yhat=ham	False negatives	True negatives

Recall

- Recall

$$\frac{\text{True positives}}{\text{True positives + False negatives}}$$

Predicted\Observed	y=spam	y=ham
yhat = spam	True positives	False positives
yhat=ham	False negatives	True negatives

Recall

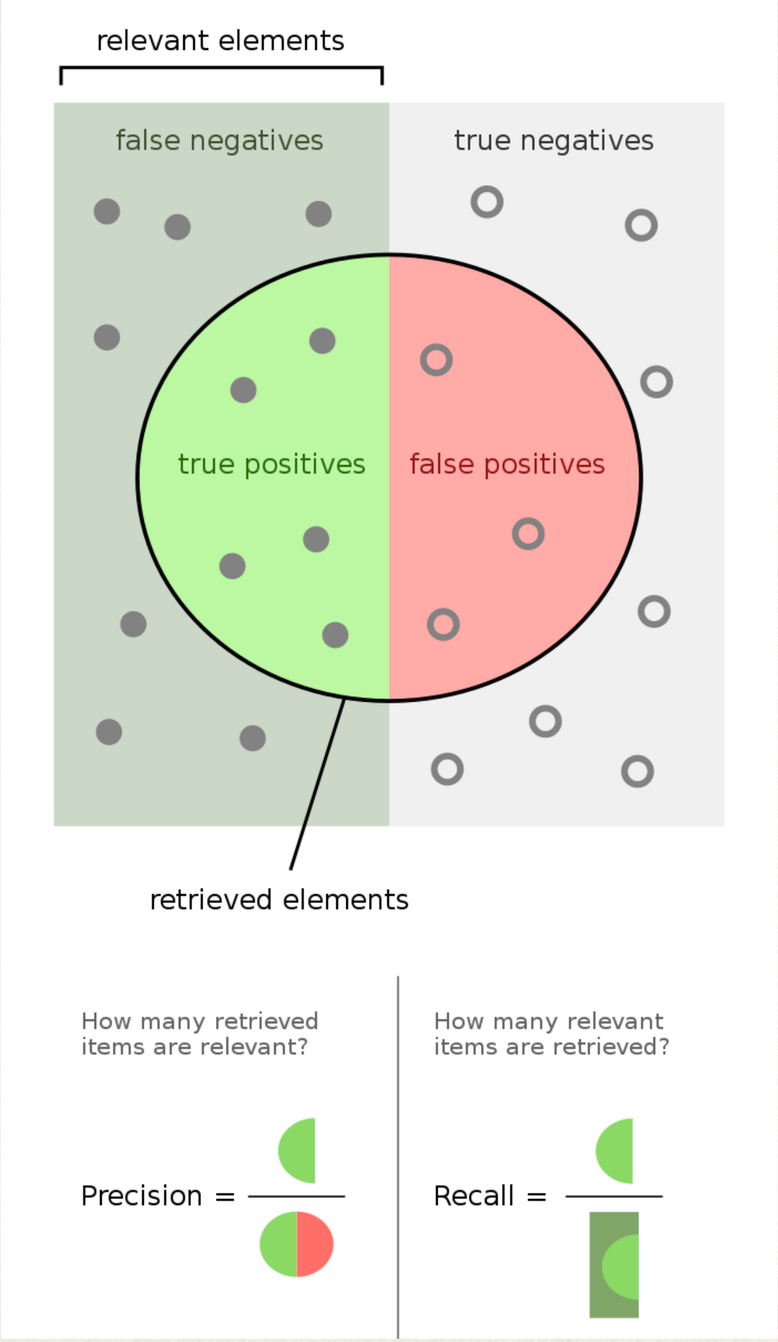
- Recall

$$\frac{\text{True positives}}{\text{True positives + False negatives}}$$

Predicted\Observed	y=spam	y=ham
yhat = spam	True positives	False positives
yhat=ham	False negatives	True negatives

- We want number of false negatives to be low and recall to be high

Precision and Recall

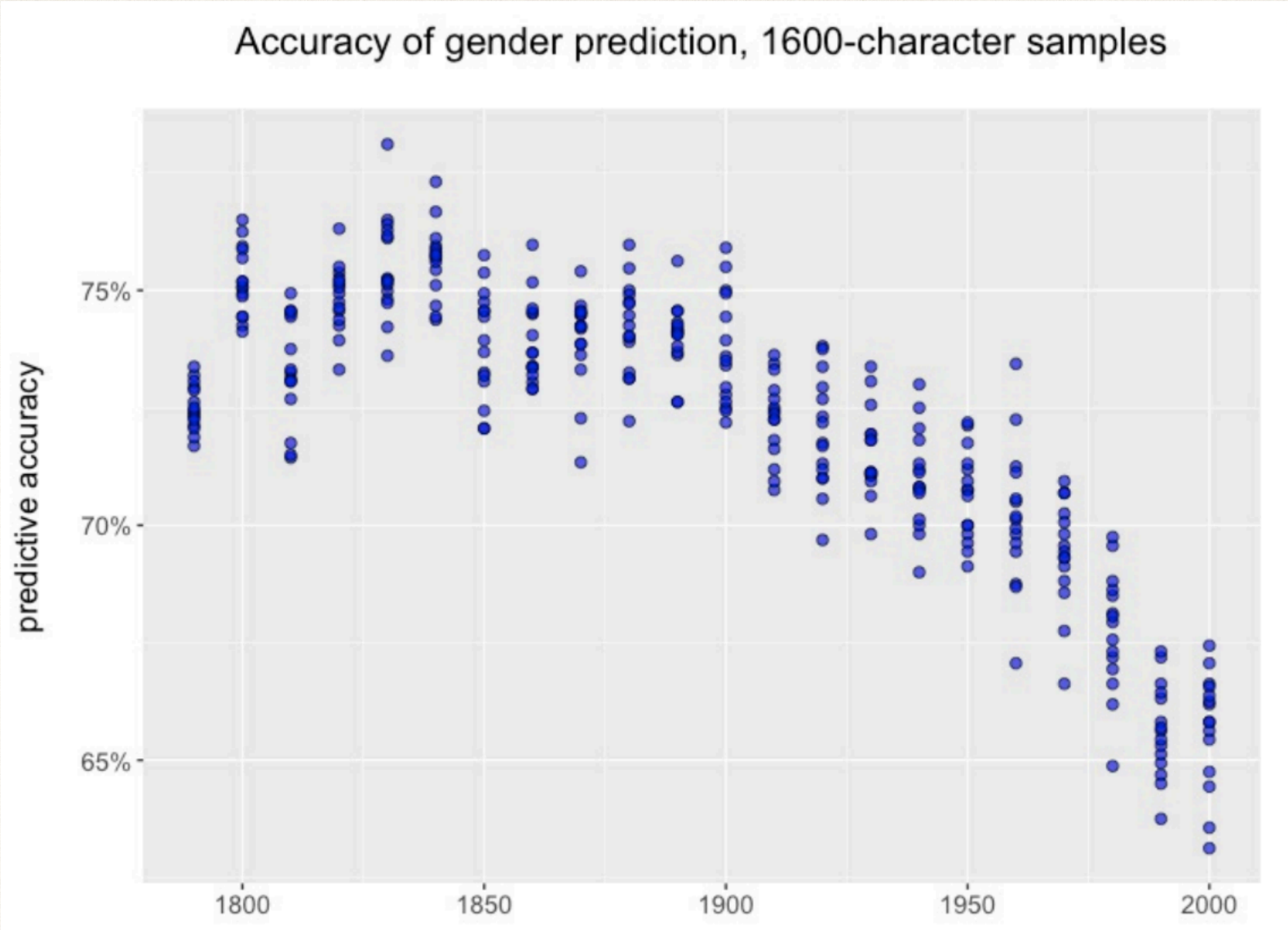


Source: Wikipedia

F1

- We can combine precision and recall into a single metric by taking the harmonic mean of the two quantities.
- $$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

How is classifier used on real world problems?



Underwood, Ted, David Bamman, and Sabrina Lee. 2018. “The Transformation of Gender in English-Language Fiction.” Journal of Cultural Analytics 3 (2). https://doi.org/10.22148/16.019.

In class

- Logistic regression demo