How to automatically label text?

Sandeep Soni

09/26/2024

Question for the day

"How to predict a label for given text?"

Agenda

- Classification problem
- Naive Bayes
- Logistic regression
- Training/test setup
- Evaluation metrics

Text classification

- Input: document (e.g., email)
- Output: label (e.g., spam/ham)

Text classification problems

Task	x	y
Language ID	text	{english, mandarin, hindi, …}
spam classification	email	{spam, ham}
party affiliation	speech	{republican, democrat}
sentiment analysis	text	{positive, negative, mixed, neutral}
music genre	lyric	{rock, pop, jazz, rap,…}

Formal task

Formal task

- x is an instance (e.g., an email)
  - $x \in \mathcal{X}$ (e.g., set of emails)

Formal task

- x is an instance (e.g., an email)
  - $x \in \mathcal{X}$ (e.g., set of emails)
- y is the desired label (e.g., spam)
  - $y \in \mathcal{Y}$ (e.g., {spam, ham})

Formal task

- x is an instance (e.g., an email)
  - $x \in \mathcal{X}$ (e.g., set of emails)
- y is the desired label (e.g., spam)
  - $y \in \mathcal{Y}$ (e.g., {spam, ham})
- $y = h(x)$
  - h maps instances to labels

Classification

- True h is unknown so find $\hat{h}$ that's a closest approximation

Rule based classification

- $\hat{h}$ ("I wish to discuss personal investment business matters with you so as to be able to learn of the available investment opportunities in your region or country.")

$\hat{h}$	if email contains phrase "investment opportunities" then spam

Supervised learning

- Learn $\hat{h}$ from training data given in the form of several $<x, y>$ pairs

$$\hat{h}(x)$$

- Still unresolved:
  - How to learn this mapping function? (e.g., which method to use, how to optimize, etc)
  - How to represent the input to this function?

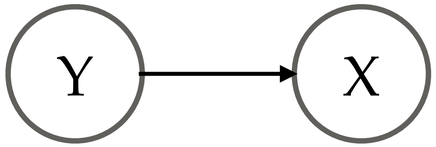
Naive Bayes

- One simple yet effective classification method is Naive Bayes
  - Similar to LDA, it's a generative model
  - We'll represent input text as a bag of words vector

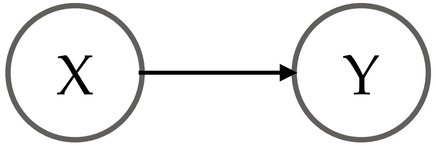
Refresher: Chain rule

- If X and Y are random variables, the joint probability can be factorized using chain rule

$$P(X, Y) = P(Y)P(X|Y)$$



$$P(X, Y) = P(X)P(Y|X)$$



Refresher: Chain rule

- If there are three variables, then chain rule gives:

$$P(x, y, z) = P(x)P(y \mid x)P(z \mid x, y)$$

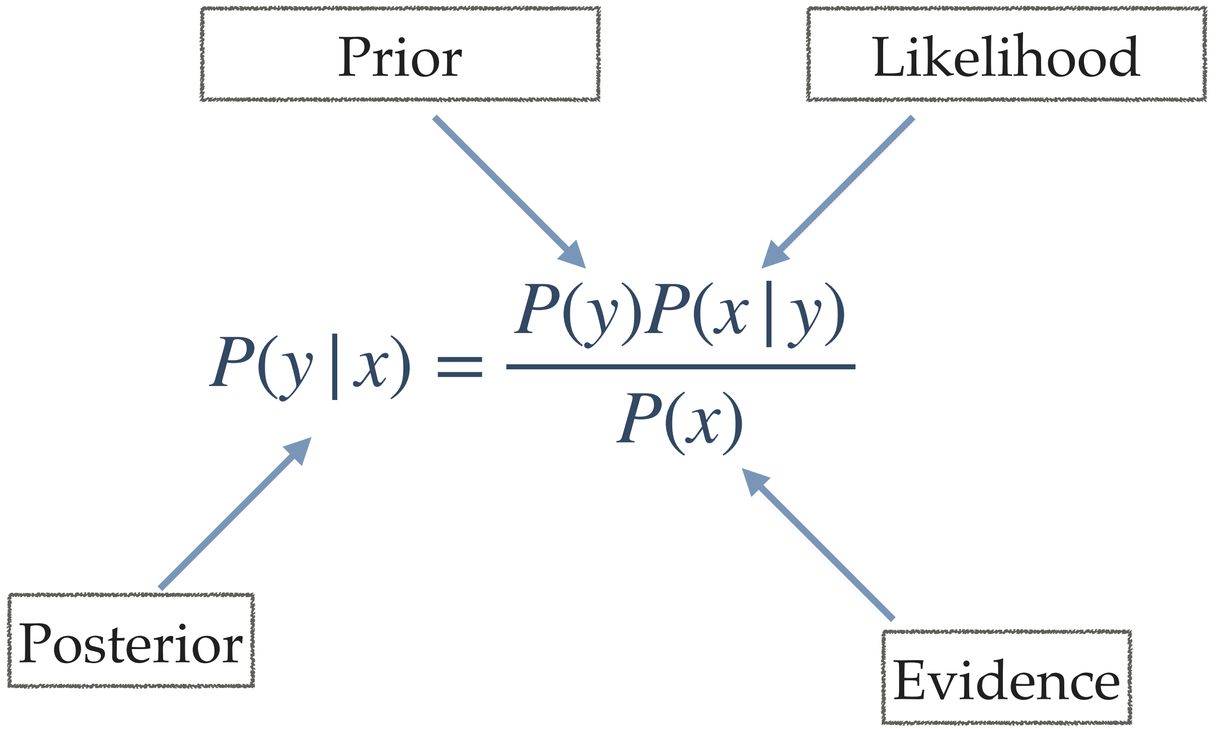
- Just like two variables, permutations are equivalent

Refresher: Chain rule

- In general, for multiple variables, chain rule is:

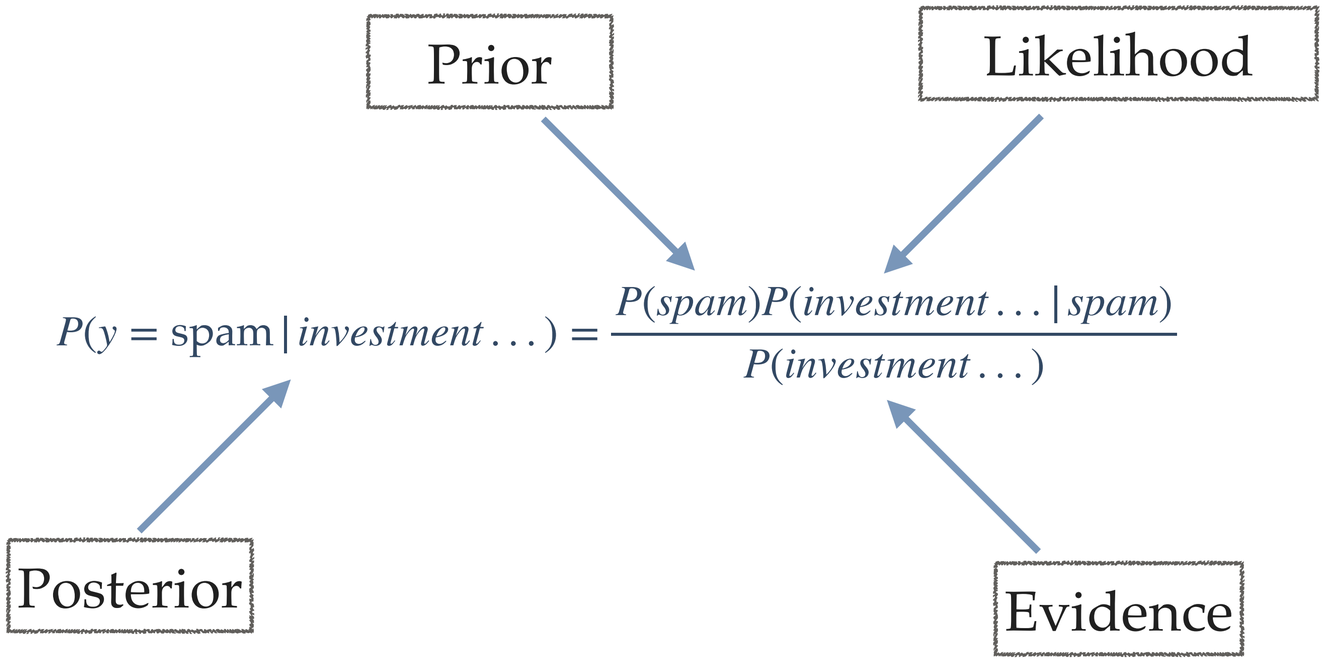
$$P(y, x_1, x_2, \ldots, x_n) = P(y)P(x_1 \mid y)P(x_2 \mid x_1, y)\ldots P(x_n \mid x_{n-1}, \cdots, y)$$

Refresher: Bayes Theorem



Refresher: Bayes Theorem

- For spam classification:
- Prior how probable is to see any email labeled as spam
- Likelihood how likely are we to see these words words in a spam email
- Evidence how probable are these words in any email
- Posterior how probable it is to label this email as spam after seeing the words in it



Refresher: Bayes Theorem

- We can express the marginal probability in the denominator as:

$$P(y \mid x) = \frac{P(y)P(x \mid y)}{P(x)}$$

$$P(y \mid x) = \frac{P(y)P(x \mid y)}{\sum_{y'} P(x, y')}$$

$$P(y \mid x) = \frac{P(y)P(x \mid y)}{\sum_{y'} P(y')P(x \mid y')}$$

Refresher: probability distributions

Y is binary RV

Y	P(Y)
TRUE	0.25
FALSE	0.75

Marginal distribution

X and Y both binary RVs

X	Y	P(X,Y)
TRUE	TRUE	0.1
TRUE	FALSE	0.3
FALSE	TRUE	0.25
FALSE	FALSE	0.35

Joint distribution

X and Y both binary RVs

P(Y | Y)	X=TRUE	X=FALSE
Y=TRUE	0.2	0.8
Y=FALSE	0.7	0.3

Conditional distribution

Refresher: probability distributions

Y is binary RV

Y	P(Y)
TRUE	0.25
FALSE	0.75

Marginal distribution

X and Y both binary RVs

X	Y	P(X,Y)
TRUE	TRUE	0.1
TRUE	FALSE	0.3
FALSE	TRUE	0.25
FALSE	FALSE	0.35

Joint distribution

X and Y both binary RVs

P(Y | Y)	X=TRUE	X=FALSE
Y=TRUE	0.2	0.8
Y=FALSE	0.7	0.3

Conditional distribution

Distribution specified by n-1 parameters if Y takes n values

Refresher: probability distributions

Y is binary RV

Y	P(Y)
TRUE	0.25
FALSE	0.75

Marginal distribution

X and Y both binary RVs

X	Y	P(X,Y)
TRUE	TRUE	0.1
TRUE	FALSE	0.3
FALSE	TRUE	0.25
FALSE	FALSE	0.35

Joint distribution

X and Y both binary RVs

P(Y | Y)	X=TRUE	X=FALSE
Y=TRUE	0.2	0.8
Y=FALSE	0.7	0.3

Conditional distribution

Distribution specified by n-1 parameters if Y takes n values

Distribution specified by 2^k - 1 parameters if there are K binary RVs

Refresher: probability distributions

Y is binary RV

Y	P(Y)
TRUE	0.25
FALSE	0.75

Marginal distribution

Distribution specified by n-1 parameters if Y takes n values

X and Y both binary RVs

X	Y	P(X,Y)
TRUE	TRUE	0.1
TRUE	FALSE	0.3
FALSE	TRUE	0.25
FALSE	FALSE	0.35

Joint distribution

Distribution specified by 2^k - 1 parameters if there are K binary RVs

X and Y both binary RVs

P(Y | Y)	X=TRUE	X=FALSE
Y=TRUE	0.2	0.8
Y=FALSE	0.7	0.3

Conditional distribution

Distribution specified by v(u-1) parameters if X takes u different values and Y takes v different values.

Towards classification

- One way to learn the mapping $\hat{h}(x)$ is to learn the posterior distribution $P(y|x)$
- x is text, so not just one random variable but a bunch of variables, so we have to estimate $P(y|x_1, x_2, x_3, \ldots, x_n)$
- Every x_i corresponds to a word or feature.

Towards classification

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

Towards classification

- Joint probability in the numerator can be factorized using chain rule

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

Continue to apply chain rule to the numerator (denominator is ignored for now)

Continue to apply chain rule to the numerator (denominator is ignored for now)

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

Continue to apply chain rule to the numerator (denominator is ignored for now)

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

Continue to apply chain rule to the numerator (denominator is ignored for now)

$$
\begin{aligned}
P(y \mid x_1, x_2, \ldots, x_n) &\propto P(y)\ P(x_1, x_2, \ldots, x_n \mid y) \\
&\propto P(y)\ P(x_1 \mid y)\ P(x_2, \ldots, x_n \mid y, x_1) \\
&\propto P(y)\ P(x_1 \mid y)\ P(x_2 \mid y, x_1)\ P(x_3, \ldots, x_n \mid y, x_1, x_2)
\end{aligned}
$$

Continue to apply chain rule to the numerator (denominator is ignored for now)

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

Continue to apply chain rule to the numerator (denominator is ignored for now)

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

This is intractable!

Naive Bayes

- Naive conditional independence assumption

$$P(x_i \mid x_{i-1}, x_{i-2}, \ldots, y) = P(x_i \mid y)$$

- Given the category, the words (features) are independent of each other
- Under naive Bayes, P(“prince”|spam) = P(“prince”|spam, “kenyan”)

Now we can rewrite the posterior probability as:

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

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

This is tractable!

Now we can rewrite the posterior probability as:

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

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

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

This is tractable!

We can estimate these probabilities by simply counting the instances in training data

$$P(y = spam) = \frac{\#\text{samples labeled spam}}{\#\text{ samples}}$$

$$P(\text{"kenyan"} \mid \text{spam}) = \frac{\#\text{samples labeled spam and contain "kenyan"}}{\#\text{samples labeled spam}}$$

Picking the label

- Once you estimate the probabilities from training data, you can pick the label that maximizes the posterior

$$P(y = \text{spam} \mid \text{text}) > P(y = \text{ham} \mid \text{text})$$ Spam

otherwise Ham

Naive Bayes

Naive Bayes

- Let’s rewrite the posterior probability under naive Bayes

Naive Bayes

- Let's rewrite the posterior probability under naive Bayes

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

Naive Bayes

- Let's rewrite the posterior probability under naive Bayes

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

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

Naive Bayes

- Let's rewrite the posterior probability under naive Bayes

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

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

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

Naive Bayes

- Let's rewrite the posterior probability under naive Bayes

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Weights

Features

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

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

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

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

Weights

Features

Naive Bayes is a linear model

Logistic Regression

Logistic Regression

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

Logistic Regression

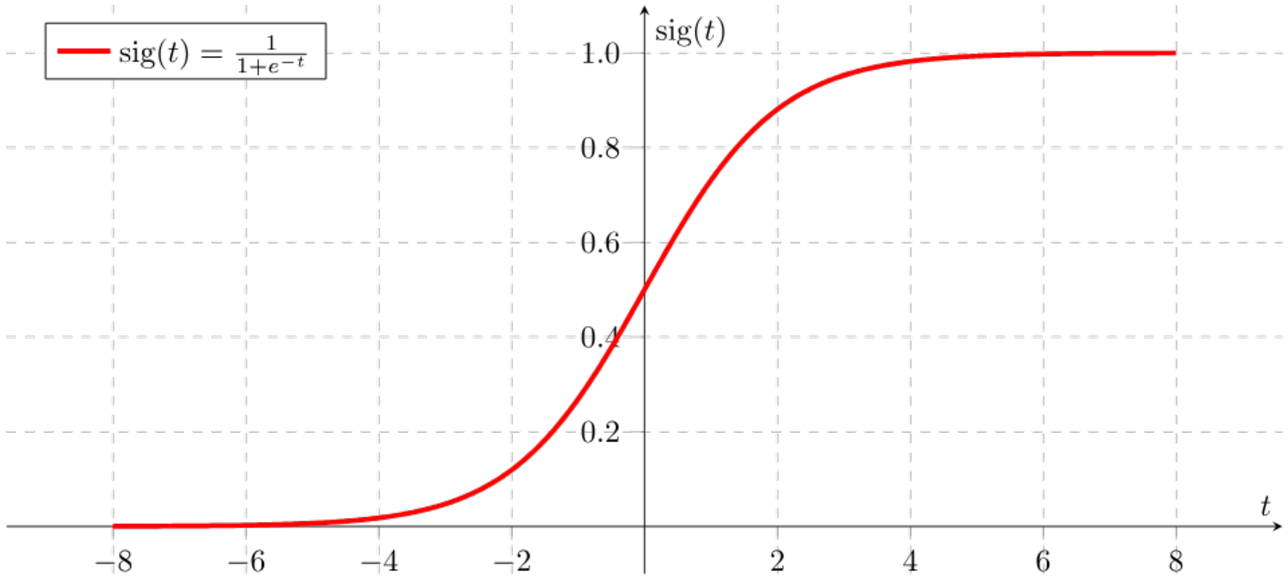
In logistic regression, we directly model the conditional probability of the label given the text

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

Logistic Regression

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

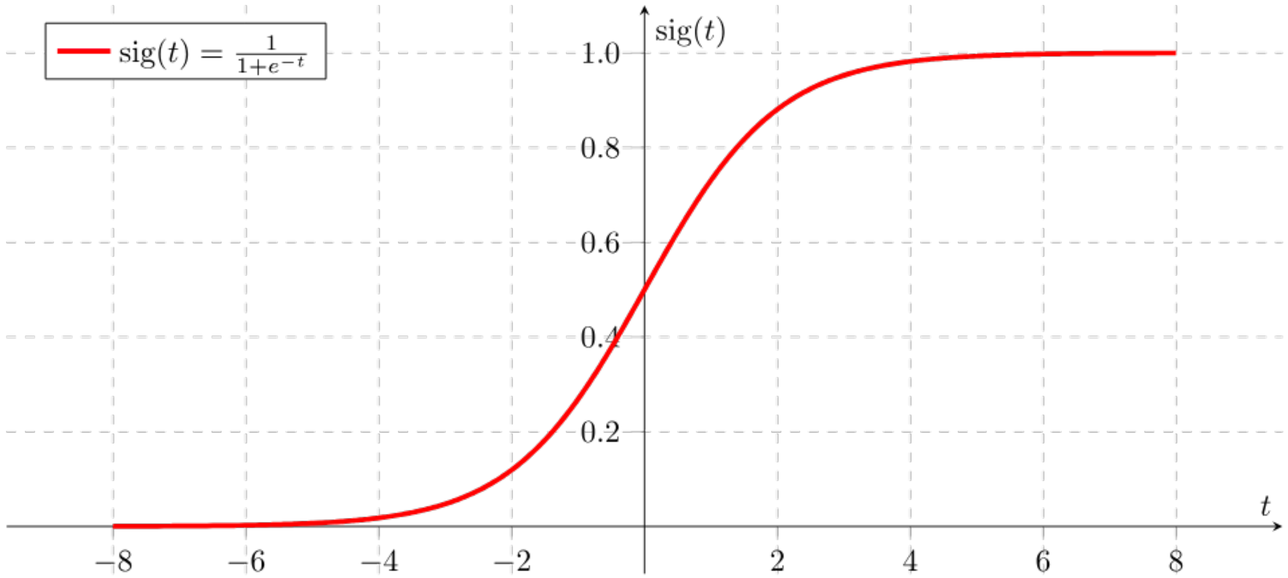
$$P(y \mid x) = \text{sig}(w^T f(x))$$



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, no closed form solution

Learning

- The objective is to learn the vector of weights w
- Unlike Naive Bayes, no closed form solution
- Weights are learned by optimizing some criterion on training data

Generalized classification

Generalized classification

- Alternatives to linear models also exist.

Generalized classification

- Alternatives to linear models also exist.
- $P(y \mid x) = \mathtt{sig}(g(f(x)))$

Generalized classification

- Alternatives to linear models also exist.
- $P(y \mid x) = \text{sig}(g(f(x)))$
- For multi-class classification, $P(y \mid x) = \text{softmax}(g(f(x)))$

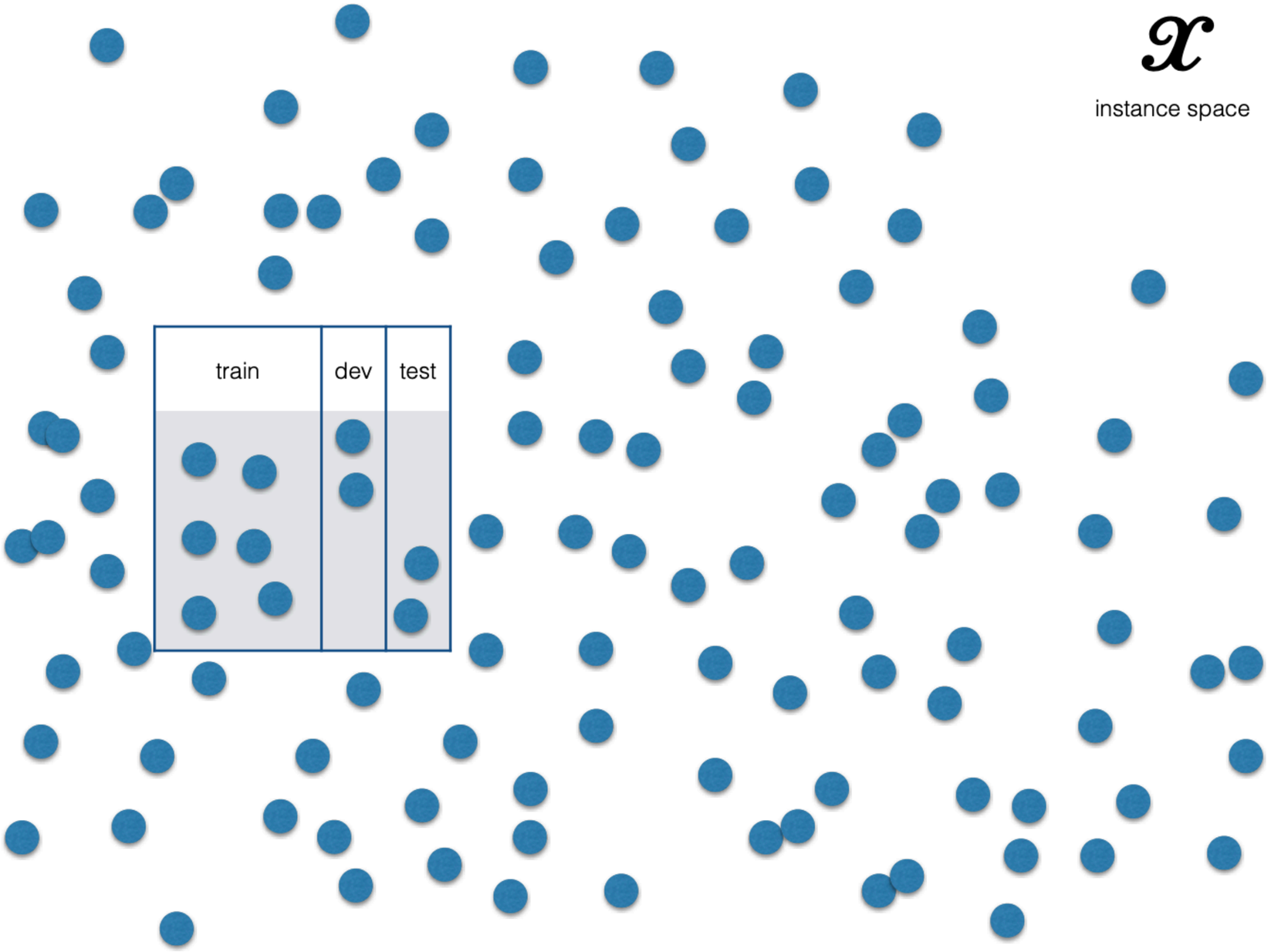
Generalized classification

- Alternatives to linear models also exist.
- $P(y|x) = \text{sig}(g(f(x)))$ g could be a neural network
- For multi-class classification, $P(y|x) = \text{softmax}(g(f(x)))$

Generalized forecasting

Generalized forecasting

- Same model can be used to predict continuous values
- $y = g(f(x))$
- If g is a linear function, this is linear regression



Slide from David Bamman

Experiment design

- Training set is to estimate parameters of the model
- Development set is to perform model selection
- Test set for evaluation

Experiment design

- Typically, we use 80% data for training, 10% for model selection and 10% for evaluation
- One should be careful never to use to development or test data to do estimation

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

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

Confusion Matrix

Confusion Matrix

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

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

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

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

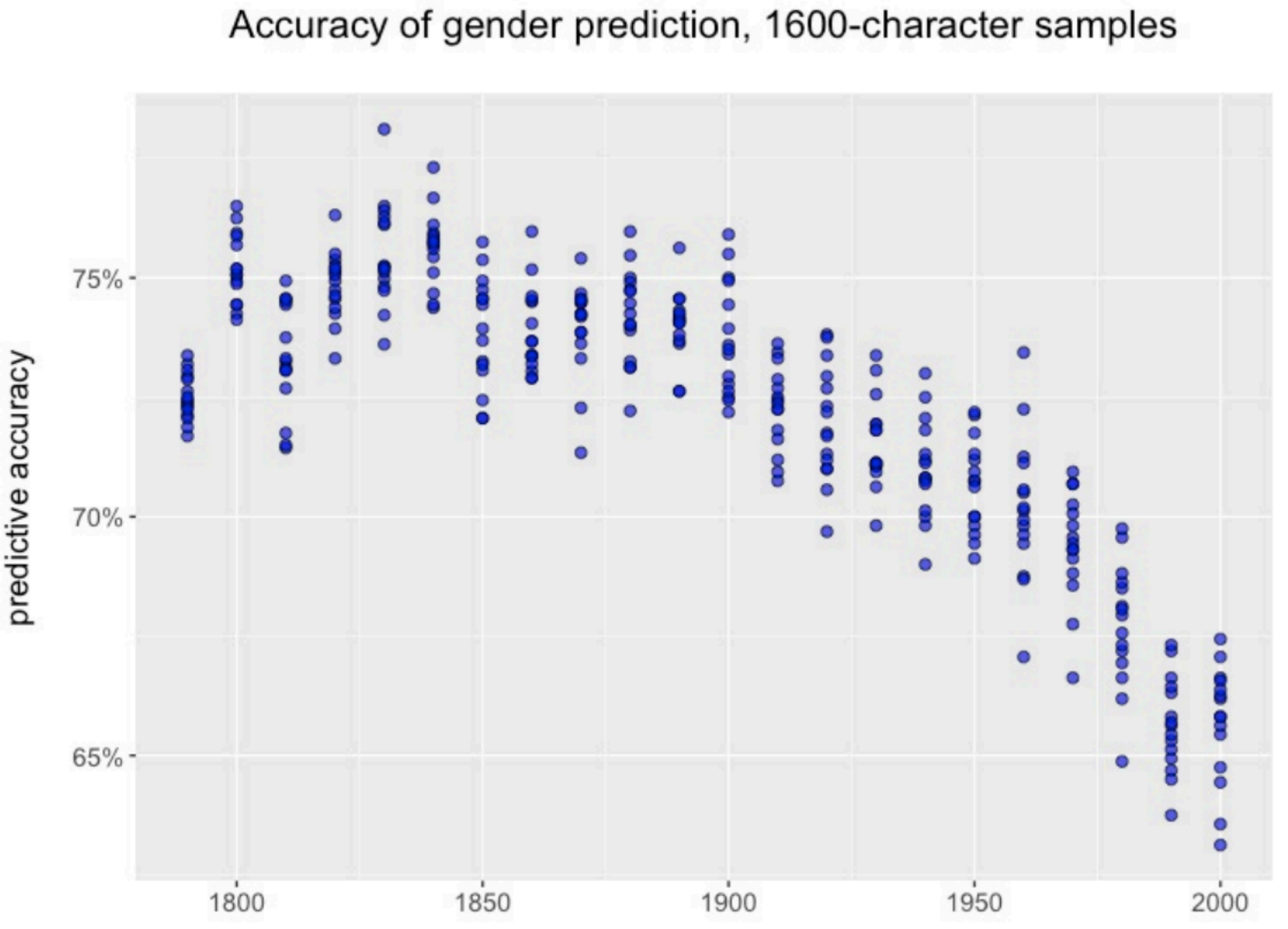
Accuracy

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

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

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

- Text classification demo