Natural Language Processing

Text classification

Yulia Tsvetkov

yuliats@cs.washington.edu

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Announcements

- We'll practice quiz set up today (last 5 min of the class)
- Course schedule and readings are updated on the website until Apr 25

Readings

- Eis 2 https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf
- J&M III 4 https://web.stanford.edu/~jurafsky/slp3/4.pdf
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In Proceedings of EMNLP, 2002
- Andrew Y. Ng and Michael I. Jordan, On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes, In Proceedings of NeurIPS, 2001.

We'll consider alternative models for classification

- Supervised text classification
  - Rule-based
  - Probabilistic
    - Generative models
      - Naïve Bayes
    - Discriminative models
      - Linear models
        - Multinomial logistic regression (aka MaxEnt)
      - Non-linear models
        - Multilayer perceptron

Generative and discriminative models

- Generative model: a model that calculates the probability of the input data itself

$$P(X, Y)$$

joint

- Discriminative model: a model that calculates the probability of a latent trait given the data

$$P(Y \mid X)$$

conditional

Generative and discriminative models

imagenet

imagenet

Generative model

- Build a model of what's in a cat image
  - Knows about whiskers, ears, eyes
  - Assigns a probability to any image:
  - how cat-y is this image?
- Also build a model for dog images

imagenet

imagenet

Now given a new image:

Run both models and see which one fits better

Discriminative model

Just try to distinguish dogs from cats

Oh look, dogs have collars! Let's ignore everything else

Generative and discriminative models

- Generative text classification: Learn a model of the joint P(X, y), and find

$$\hat{y} = \underset{\tilde{y}}{\operatorname{argmax}}\ P(X, \tilde{y})$$

- Discriminative text classification: Learn a model of the conditional P(y | X), and find

$$\hat{y} = \underset{\tilde{y}}{\operatorname{argmax}}\ P(\tilde{y}|X)$$

Andrew Y. Ng and Michael I. Jordan, On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes, Advances in Neural Information Processing Systems 14 (NIPS), 2001.

Finding the correct class c from a document d in Generative vs Discriminative Classifiers

- Naive Bayes

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \overbrace{P(d|c)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}}$$

- Logistic Regression

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \overset{\text{posterior}}{P(c|d)}$$

We'll consider alternative models for classification

- Supervised text classification
  - Rule-based
  - Probabilistic
    - Generative models
      - Naïve Bayes
    - Discriminative models
      - Linear models
        - Multinomial logistic regression (aka MaxEnt)
      - Non-linear models
        - Multilayer perceptron

Generative text classification: Naïve Bayes

$$\mathrm{C}_{NB} = \underset{c}{\text{argmax}} P(c|d) = \underset{c}{\text{argmax}} \frac{P(d|c)P(c)}{P(d)} \propto$$

Bayes rule

$$\underset{c}{\text{argmax}} P(d|c)P(c) =$$

same denominator

$$\underset{c}{\text{argmax}} P(w_1, w_2, \ldots, w_n|c)P(c) =$$

representation

$$\underset{c_j}{\text{argmax}} \ P(c_j) \prod_i P(w_i|c)$$

conditional independence

Underflow prevention: log space

- Multiplying lots of probabilities can result in floating-point underflow
- Since log(xy) = log(x) + log(y)
  - better to sum logs of probabilities instead of multiplying probabilities
- Class with highest un-normalized log probability score is still most probable

$$C_{NB} = \underset{c_j}{\text{argmax}}\ P(c_j) \prod_i P(w_i|c)$$

$$C_{NB} = \underset{c_j}{\text{argmax}}\ log(P(c_j)) + \sum_i log(P(w_i|c))$$

- Model is now just max of sum of weights

Learning the multinomial naïve Bayes

- How do we learn (train) the NB model?

Learning the multinomial naïve Bayes

- How do we learn (train) the NB model?
- We learn $P(c)$ and $P(w_i|c)$ from training (labeled) data

$$C_{NB} = \underset{c_j}{\text{argmax}}\ log(P(c_j)) + \sum_i log(P(w_i|c))$$

Parameter estimation

- Parameter estimation during training
- Concatenate all documents with category c into one mega-document
- Use the frequency of w_i in the mega-document to estimate the word probability

$$C_{NB} = \underset{c_j}{\operatorname{argmax}}\ log(P(c_j)) + \sum_i log(P(w_i|c))$$

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Parameter estimation

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

- fraction of times word w_i appears among all words in documents of topic c_j

- Create mega-document for topic j by concatenating all docs in this topic
  - Use frequency of w in mega-document

Problem with Maximum Likelihood

- What if we have seen no training documents with the word "fantastic" and classified in the topic positive?

Problem with Maximum Likelihood

- What if we have seen no training documents with the word “fantastic” and classified in the topic positive?

$$\hat{P}(\text{“}fantastic\text{”}|c = \text{positive}) = \frac{count(\text{“}fantastic\text{”}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

- Zero probabilities cannot be conditioned away, no matter the other evidence!

$$\underset{c_j}{\text{argmax}}\ P(c_j) \prod_i P(w_i|c)$$

Laplace (add-1) smoothing for naïve Bayes

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V}(count(w, c_j) + 1)}$$

Laplace (add-1) smoothing for naïve Bayes

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V}(count(w, c_j) + 1)}$$

$$= \frac{count(w_i, c_j) + 1}{\left(\sum_{w \in V} count(w, c_j)\right) + |V|}$$

Example

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Example

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(c) = \frac{N_c}{N}$$

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Example

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

Example

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7
P(Tokyo|*c*) = (0+1) / (8+6) = 1/14
P(Japan|*c*) = (0+1) / (8+6) = 1/14
P(Chinese|*j*) = (1+1) / (3+6) = 2/9
P(Tokyo|*j*) = (1+1) / (3+6) = 2/9
P(Japan|*j*) = (1+1) / (3+6) = 2/9

Choosing a class:

$P(c|d5) \propto 3/4 * (3/7)^3 * 1/14 * 1/14 \approx 0.0003$

$P(j|d5) \propto 1/4 * (2/9)^3 * 2/9 * 2/9 \approx 0.0001$

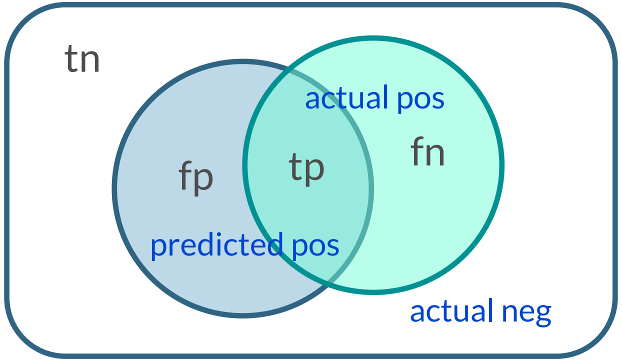
Summary: naïve Bayes is not so naïve

- Naïve Bayes is a probabilistic model
- Naïve because is assumes features are independent of each other for a class
- Very fast, low storage requirements
- Robust to Irrelevant Features
- Very good in domains with many equally important features
- Optimal if the independence assumptions hold: If assumed independence is correct, then it is the Bayes Optimal Classifier for problem
- A good dependable baseline for text classification
    - But we will see other classifiers that give better accuracy

Classification evaluation

- Contingency table: model's predictions are compared to the correct results
  - a.k.a. confusion matrix

	actual pos	actual neg
predicted pos	true positive (tp)	false positive (fp)
predicted neg	false negative (fn)	true negative (tn)



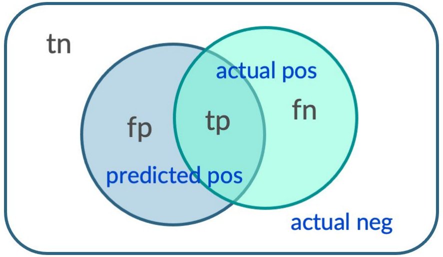
Classification evaluation

- Borrowing from Information Retrieval, empirical NLP systems are usually evaluated using the notions of precision and recall

Classification evaluation

- Precision (P) is the proportion of the selected items that the system got right in the case of text categorization
  - it is the % of documents classified as “positive” by the system which are indeed “positive” documents
- Reported per class or average

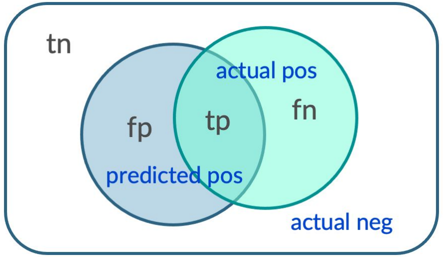
$$\text{precision} = \frac{true\ positives}{true\ positives + false\ positives} = \frac{tp}{tp + fp}$$



Classification evaluation

- Recall (R) is the proportion of actual items that the system selected in the case of text categorization
  - it is the % of the “positive” documents which were actually classified as “positive” by the system
- Reported per class or average

$$\text{recall} = \frac{\mathit{true\ positives}}{\mathit{true\ positives} + \mathit{false\ negatives}} = \frac{tp}{tp + fn}$$



Classification evaluation

- We often want to trade-off precision and recall
  - typically: the higher the precision the lower the recall
  - can be plotted in a precision-recall curve
- It is convenient to combine P and R into a single measure
  - one possible way to do that is F measure

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R} \qquad \text{for } \beta=1,\ F_1 = \frac{2PR}{P+R}$$

Classification evaluation

- Additional measures of performance: accuracy and error
  - accuracy is the proportion of items the system got right
  - error is its complement

	actual pos	actual neg
predicted pos	true positive (tp)	false positive (fp)
predicted neg	false negative (fn)	true negative (tn)

$$\text{accuracy} = \frac{tp+tn}{tp+fp+tn+fn}$$

Micro- vs. macro-averaging

If we have more than one class, how do we combine multiple performance measures into one quantity?

- Macroaveraging
  - Compute performance for each class, then average.
- Microaveraging
  - Collect decisions for all classes, compute contingency table, evaluate.

Logistic regression

- Supervised text classification
  - Rule-based
  - Probabilistic
    - Generative models
      - Naïve Bayes
    - Discriminative models
      - Linear models
        - Multinomial logistic regression (aka MaxEnt)
      - Non-linear models
        - Multilayer perceptron

Logistic regression classifier

- Important analytic tool in natural and social sciences
- Baseline supervised machine learning tool for classification
- Is also the foundation of neural networks

Text classification

Input:

- a document d (e.g., a movie review)
- a fixed set of classes $C = \{c_1, c_2, \ldots c_j\}$ (e.g., positive, negative, neutral)

Output

- a predicted class $\hat{y} \in C$

Binary classification in logistic regression

- Given a series of input/output pairs:
  - $(x^{(i)}, y^{(i)})$

- For each observation $x^{(i)}$
  - We represent $x^{(i)}$ by a feature vector $\{x_1, x_2, \ldots, x_n\}$
  - We compute an output: a predicted class $\hat{y}^{(i)} \in \{0,1\}$

Features in logistic regression

- For feature $x_i \in \{x_1, x_2, \ldots, x_n\}$, weight $w_i \in \{w_1, w_2, \ldots, w_n\}$ tells us how important is x_i
  - x_i = "review contains 'awesome'": w_i = +10
  - x_j = "review contains horrible": w_j = -10
  - x_k = "review contains 'mediocre'": w_k = -2

Logistic Regression for one observation x

- Input observation: vector $x^{(i)} = \{x_1, x_2, \ldots, x_n\}$
- Weights: one per feature: $W = [w_1, w_2, \ldots, w_n]$
  - Sometimes we call the weights $\theta = [\theta_1, \theta_2, \ldots, \theta_n]$
- Output: a predicted class $\hat{y}^{(i)} \in \{0,1\}$

multinomial logistic regression: $\hat{y}^{(i)} \in \{0,1, 2, 3, 4\}$

How to do classification

- For each feature x_i, weight w_i tells us importance of x_i
  - (Plus we'll have a bias b)
  - We'll sum up all the weighted features and the bias

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

$$z = w \cdot x + b$$

If this sum is high, we say $y=1$; if low, then $y=0$

But we want a probabilistic classifier

We need to formalize “sum is high”

- We’d like a principled classifier that gives us a probability, just like Naive Bayes did

- We want a model that can tell us:
  - $p(y=1|x;\theta)$
  - $p(y=0|x;\theta)$

The problem: z isn't a probability, it's just a number!

- z ranges from $-\infty$ to ∞

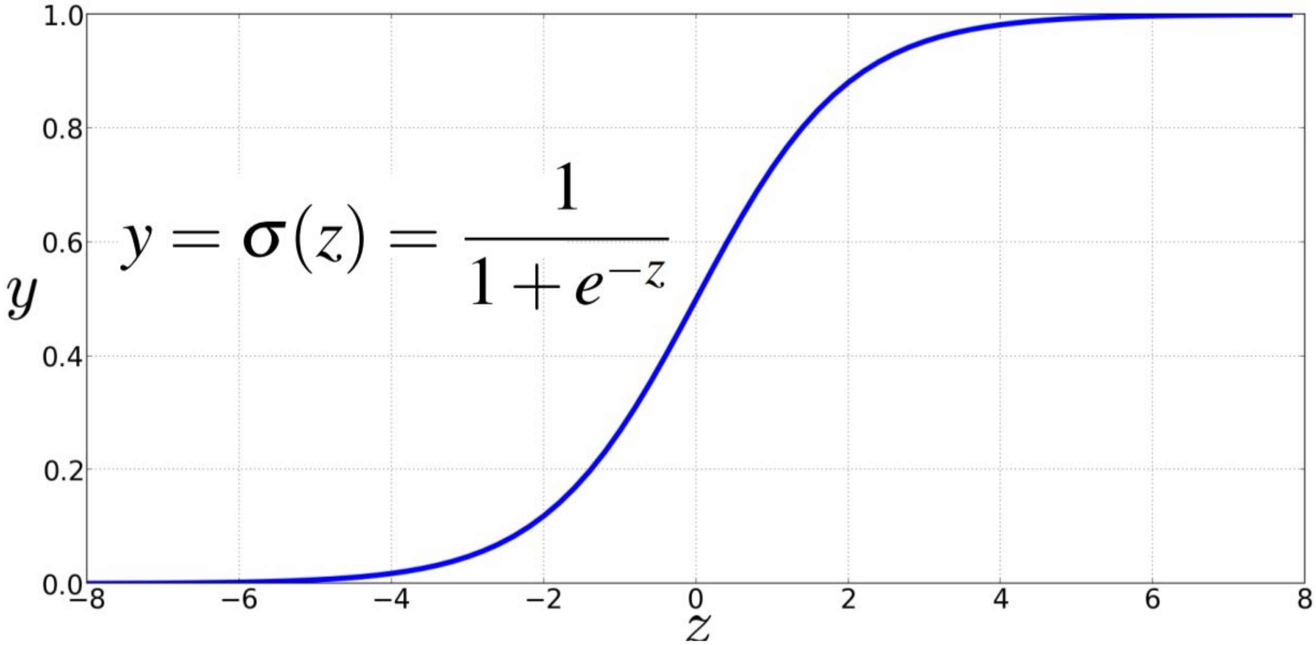
$$z = w \cdot x + b$$

- Solution: use a function of z that goes from 0 to 1

"sigmoid" or "logistic" function

$$y = \sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+\exp(-z)}$$

The very useful sigmoid or logistic function



Idea of logistic regression

- We'll compute w·x+b
- And then we'll pass it through the sigmoid function:

$$\sigma(w \cdot x + b)$$

- And we'll just treat it as a probability

Making probabilities with sigmoids

$$
\begin{aligned}
P(y=1) &= \sigma(w \cdot x + b) \\
&= \frac{1}{1+\exp(-(w \cdot x + b))}
\end{aligned}
$$

Making probabilities with sigmoids

$$
\begin{aligned}
P(y=1) &= \sigma(w \cdot x + b) \\
&= \frac{1}{1+\exp(-(w \cdot x + b))}
\end{aligned}
$$

$$
\begin{aligned}
P(y=0) &= 1 - \sigma(w \cdot x + b) \\
&= 1 - \frac{1}{1+\exp(-(w \cdot x + b))} \\
&= \frac{\exp(-(w \cdot x + b))}{1+\exp(-(w \cdot x + b))}
\end{aligned}
$$

By the way:

$$
\begin{aligned}
P(y=0) &= 1-\sigma(w\cdot x+b) \\
&= 1-\frac{1}{1+\exp\left(-(w\cdot x+b)\right)} \\
&= \frac{\exp\left(-(w\cdot x+b)\right)}{1+\exp\left(-(w\cdot x+b)\right)}
\end{aligned}
$$

$= \sigma(-(w\cdot x+b))$

Because

$$1-\sigma(x)=\sigma(-x)$$

Turning a probability into a classifier

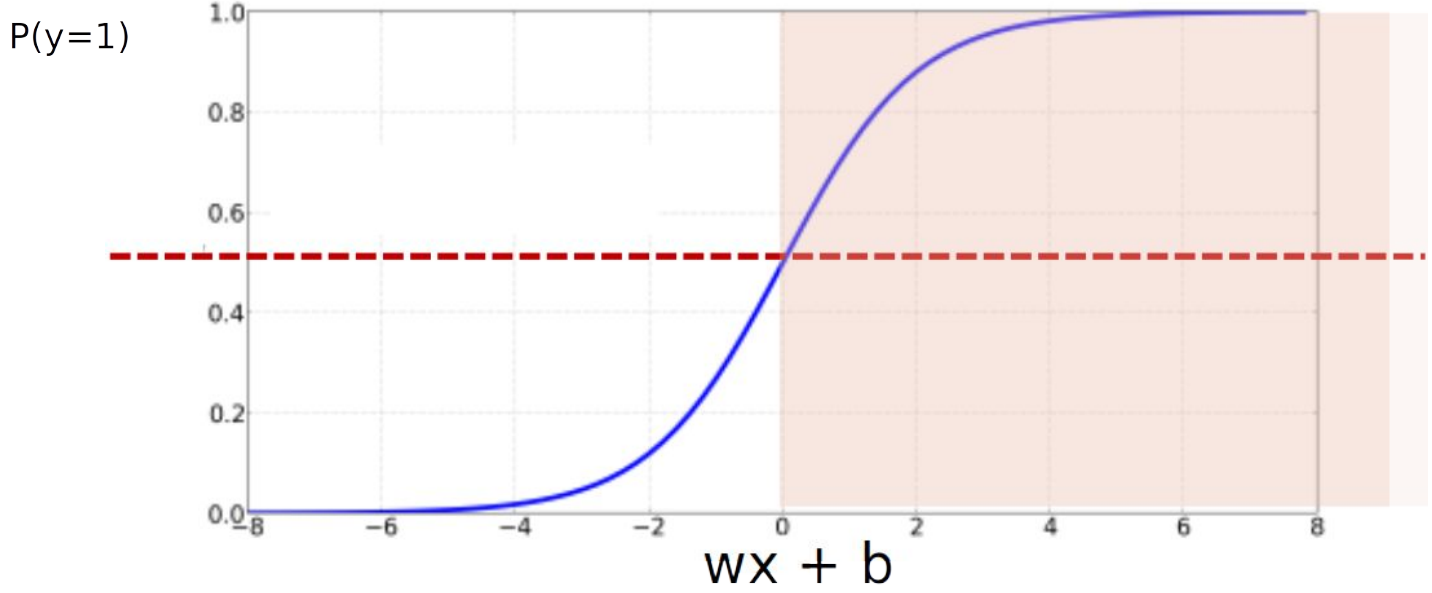
$$\hat{y} = \begin{cases} 1 & \text{if } P(y=1|x) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

- 0.5 here is called the **decision boundary**

The probabilistic classifier

$$P(y=1) = \sigma(w \cdot x + b)$$

$$= \frac{1}{1+\exp(-(w \cdot x + b))}$$



Turning a probability into a classifier

$$\hat{y} = \begin{cases} 1 & \text{if } P(y=1|x) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

if $w \cdot x + b > 0$

if $w \cdot x + b \leq 0$

Wait, where did the W's come from?

- Supervised classification:
  - A training time we know the correct label y (either 0 or 1) for each x.
  - But what the system produces at inference time is an estimate $\hat{y}$
- We want to set w and b to minimize the **distance** between our estimate $\hat{y}^{(i)}$ and the true $y^{(i)}$
  - We need a distance estimator: a **loss function** or a cost function
  - We need an **optimization algorithm** to update w and b to minimize the loss

Learning components

A loss function:

- **cross-entropy loss**

An optimization algorithm:

- **stochastic gradient descent**

Loss function: the distance between ŷ and y

We want to know how far is the classifier output $\hat{y} = \sigma(w \cdot x + b)$

from the true output: y [= either 0 or 1]

We'll call this difference: $L(\hat{y}, y)$ = how much ŷ differs from the true y

Intuition of negative log likelihood loss = cross-entropy loss

A case of conditional maximum likelihood estimation

We choose the parameters w,b that maximize

- the log probability
- of the true y labels in the training data
- given the observations x

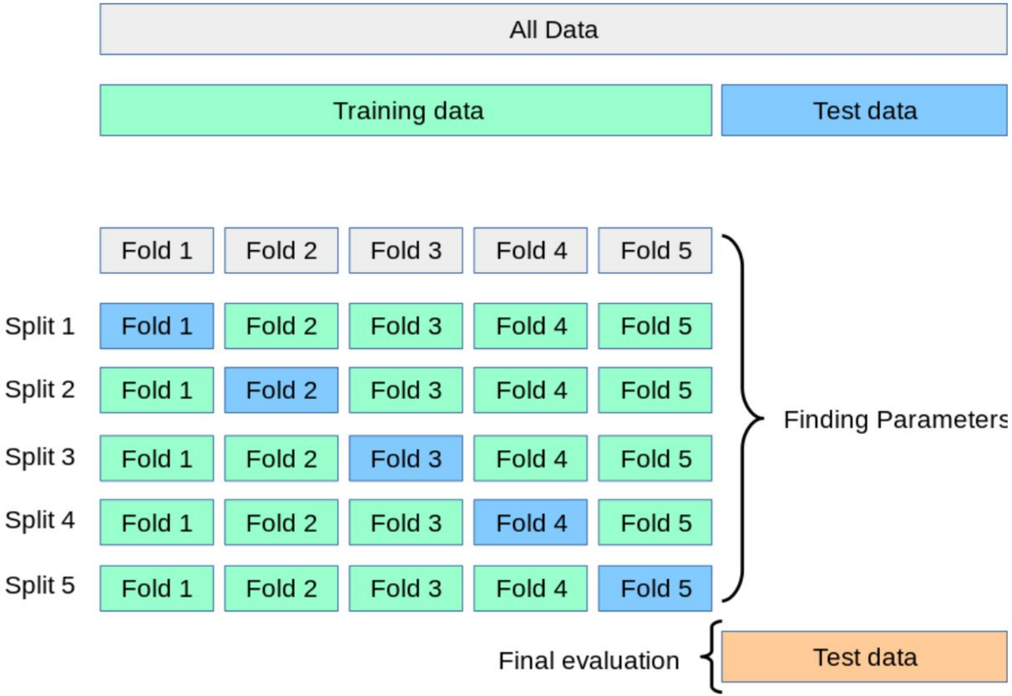
Next class:

- Deriving cross-entropy loss
- Learning weights with Stochastic Gradient Descent
- Multinomial LR

Classification common practices

- Divide the training data into k folds (e.g., $k=10$)
- Repeat k times: train on $k-1$ folds and test on the holdout fold, cyclically
- Average over the k folds' results

K-fold cross-validation



K-fold cross-validation

- Metric: P/R/F1 or Accuracy
- Unseen test set
  - avoid overfitting ('tuning to the test set')
  - more conservative estimate of performance
- Cross-validation over multiple splits
    - Handles sampling errors from different datasets
  - Pool results over each split
  - Compute pooled dev set performance