Machine learning and naive Bayes classification units

- Machine learning boot camp (last week Friday)
- Bag-of-words model **(Today)**
- Naive Bayes classification (Wednesday)
- Lab: NBC (Friday)
- Finish NBC (next week Monday)

Today:

- Finishing ML basics
- Bag-of-words model
  - Vectors as abstract views
  - Bag-of-words definition
  - Variations and options

Machine learning is a form of applied statistics with emphasis on the use of computers to statistically estimate complicated functions.

Goodfellow et al., Deep Learning, 2016. Pg 95.

Machine learning is the science (and art) of programming computers so they can learn from data. [In 1959, Arthur Samuel defined machine learning as the] field of study that gives computers the ability to learn without being explicitly programmed.

Géron, Hands-On Machine Learning, 2019. Pg 2.

[Machine learning is] a set of methods that can automatically detect patterns in data and then use the uncovered patterns to predict future data or to perform other kinds of decision-making under uncertainty.

Murphy, Machine Learning: A Probabilistic Perspective, 2012. Pg 1.

Train a function whose inputs are real-valued vectors of length D:

$$f : \mathbb{R}^D \to T$$

where T is the output type/return type/codomain/target type.

Machine learning main tasks:

- Classification, where the target type is a finite set
  - Binary classification, where $|T| = 2$
    the target is $\{\bot, \top\}$ (or $\{0, 1\}$ or $\{-1, 1\}$ ...)
- Density estimation, where the target type is $[0, 1] = \{y \in \mathbb{R} \mid 0 \leq y \leq 1\}$.
- Regression, where the target type is $\mathbb{R}$

Examples of classification:

- Given the intensity (and/or color) of pixels in an image of a hand-written digit, what numeral is this?
- Given various measurements (such as, from diagnostic imaging), is this growth malignant or benign?
- Given various measurements, what species is this plant?
- What genre (or topic) is this text?
- Is this review/comment/tweet/blog post positive or negative in sentiment? Is it real or fake?
- Is this email spam or not?
- Who is the author of this text?

Examples of regression:

- Given a year's change in employment, what is the range in that year's GDP?
- Given a car's weight and volume, what is its carbon dioxide output?
- Given various data on an real estate property, what is that property's market value?
- Given a text, what is its reading level (*Lexile*)?

- Training algorithm; model; model family; parameters

  Parameters are used to select a model from a model family. A training algorithm finds parameters.

- Training set; test set; held-out set; development (dev) set.

  The dataset available is split into the training set and the test set. The training set is used to train the model, that is, find parameters. The test set is used to evaluate the model, and, for the integrity of that evaluation, it should not be used as part of the training set. Other parts of the data can also be "held out"—that is, not used in training. For example, a development set can be used for tuning a model.

- Overfitting; generalization; regularization

  One hazard in machine learning is that the model picks up on the peculiarities of the observations in the training set instead of trends of the phenomenon that the training set is sampled from. We can detect this when a model performs well on its training set but not on the test set. We say that the model has overfit to the training data. The opposite—what we want—is for the model to generalize to the phenomenon the training data come from. Regularization is a modification made to the model family with the intent to reduce overfitting. Regularization usually takes the form of penalizing model complexity.

Coming up:

- Do CKY parsing programming assignment (Mon, Oct 27)
- Take ML basics and bag-of-words quiz (Tues, Oct 28)
- Read J&M B.(0-8, 10). (Fri, Oct 31)