Machine learning and naive Bayes classification units

- Machine learning boot camp (last week Friday)
- Finishing basic ML terms; bag-of-words model **(Today)**
- Naive Bayes classification (next week Monday)
- Lab: NBC (next week Wednesday)

Today:

- Finishing ML basics
  - Tasks and terminology
  - Models
  - The nature of data
- Bag-of-words model
  - Vectors as abstract views
  - Bag-of-words definition
  - Variations and options

Coming up:

- Take ML basics quiz (Tues, Oct 24)
- Do bag-of-words programming assignment (Wed, Oct 25)
- Read J&M 4.(0-8, 10) (Wed, Oct 25)
- Take NBC quiz (Thurs, Oct 27)
- Do NBC programming assignment (Fri, Nov 3)

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.

Machine learning main tasks:

- Regression, where the target type is $\mathbb{R}$
- Classification, where the target type is a finite set
  - Binary classification, where the target is $\{F, T\}$ (or $\{0, 1\}$ or $\{-1, 1\}$ ...)
- Density estimation, where the target type is $[0, 1]$.

Other machine learning tasks (see Goodfellow, *Deep Learning*, pg 96–100):

- Transcription, where the observations are unstructured and the targets are text.
- Machine translation, where the observations and targets are text.
- Anomaly detection, where the targets are indicators of whether the observation is atypical.
- Synthesis and sampling, where there are no observations in deployment, but rather the program produces new observations similar to those in training.
- Denoising, where the targets are corrected versions of the observations.

Iris data set:

Data has 150 instances, 4 features (sepal length, sepal width, petal length, petal width), and three target values (Setosa, Versicolour, and Virginica)

Breast Cancer Wisconsin Data set :

Data has 569 instances, 30 features (based on radius, perimeter, area, concavity, etc.), and two target values (malignant, benign)

Coming up:

- Take ML basics quiz (Tues, Oct 24)
- Do bag-of-words programming assignment (Wed, Oct 25)
- Read J&M 4.(0-8, 10) (Wed, Oct 25)
- Take NBC quiz (Thurs, Oct 27)
- Do NBC programming assignment (Fri, Nov 3)