DATA 607 Statistical and Machine Learning
Session 1

Matthew Greenberg

Department of Mathematics and Statistics
University of Calgary

Session 1 — 24.02.2020

This Evening's Agenda

Course Information

- **Instuctor:** Matthew Greenberg

  Email: `mgreenbe@ucalgary.ca`

  Office hours: Mondays and Wednesdays, 16:00-17:00

- **TA:** Mingchen Ren

  Email: `ming.ren@ucalgary.ca`

  Office hours: TBA

- **Textbook:** Aurélien Geron, *Hands on Machine Learning with Scikit-Learn and Tensorflow, 2nd Edition*, O'Reilly.

  Also useful: James et al., *An introduction to Statistical Learning (with Applications in R)*, Springer, FREE ONLINE.

- **Evaluation:** 5 homework assignments, equally weighted

  Due dates: 04.03, 11.03, 18.03, 25.03, 01.04 at 23:59

  Distribution: ~1 week before due date

  Submission: Jupyter notebook format (`.ipynb`), via D2L

  Conversion: Minimum % required for...

A+	A	A−	B+	B	B−	C+	C	C−	D+	D
95	90	85	80	75	70	65	60	55	50	45

- **Description:** From the calendar:

  Advancement of the linear statistical model including introduction to data transformation methods, classification, model assessment and selection. Exposure to both supervised learning and unsupervised learning.

- **Topics**

  1. Introduction: Models (1 session)
  2. Nonparametric models for supervised learning (3 sessions)
  3. Deep models for supervised learning (4 sessions)
  4. Unsupervised and self-supervised learning (4 sessions)

- **Software:**
  - Python: numpy, pandas, matplotlib, scikit-learn, ntlk, tensorflow

    Please ensure you have the latest stable versions of Python 3 and of the a libraries intalled (e.g., via the Anaconda platform).
  - Jupyter notebooks: localhost, Google Colab

    Please ensure you have the latest stable version of Jupyter Noteboook/JupyterLab intalled.
  - Other: Git, markdown, LaTeX

Big Picture

"Definitions"

- ***Machine learning:*** Using algorithms to learn from data.
- ***Algorithm:*** A sequence of explicit instructions for performing a computation.
- ***Learn:*** Improve a performance metric. Solve a problem.
- ***Data:*** Information. Input to an algorithm.

Jargon

What's the difference?

- Data science
- Machine learning
- Artificial intelligence
- Statistical learning
- Statistics

An answer to this question would require precise, broadly accepted definitions of these terms.

These terms suggest different points of view on similar problems and subject matter.

My point of view:

- *Data science* is characterized by its breadth and inclusivity. Exploratory analysis, visualization, and communication are core components.
- *Machine learning* and *artificial intelligence* emphasize algorithms, computation, and scale. Prediction and generalization are key performance metrics. [This course]
- *Statistical learning* emphasizes theoretical guarantees regarding consistency, rates of connvergence, etc.
- *Statistics* emphasizes sampling, experimental design, inference, confidence intervals, hypothesis tests, p-values...

Models

A *model* is a structure or a family of structures ostensibly[1] describing a data set.

A *statistical model* is a model in which this family of structures consists of *probability distributions*.

The failure of a family member to reflect the structure of data set is measured by a *loss function*.

Fitting a model is the process of determining which member of this family best describes the data set, i.e., minimizes the loss function.

Some models, after having been fit, can be used for *prediction*.

[1] apparently or purportedly, but perhaps not actually (from `lexico.com`)

Parametric and Nonparametric Models

Parameters are quantities associated to a model that must be learned from data.

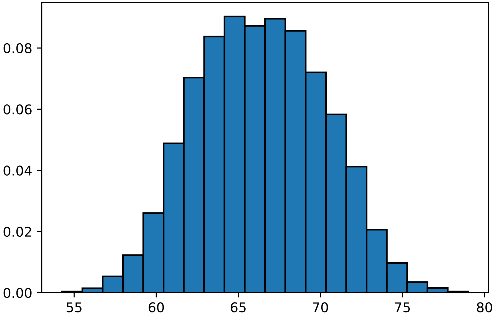
Parametric model: The number of parameters is independent of the size of the data set.

- Example: The family of Gaussian distributions $N(\mu, \sigma)$ with densities
$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)/2\sigma^2}.$$
- Example: The simple linear regression model:
$$y = \beta_0 + \beta_1 x + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

Nonparametric model: The number of parameters depends on the size of the data set.

- Example: Histogram density estimators:

$$p(x|c_1, \ldots, c_n, n) = c_i \text{ if } x \text{ is in bin } i$$



Typically, we use larger n for larger data sets.

Generative and Discrimintive Models

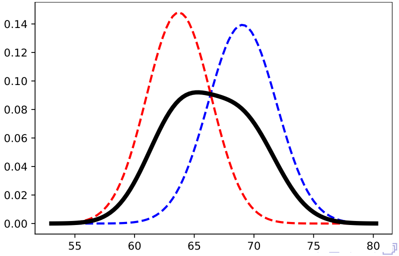
Generative model: A model of the distribution from which your data set was drawn.

- Example: Bayes classifier

$$p(x,z) = p(x|z)p(z)$$

- Mixture of Gaussians

$$p(x) = \sum_i \pi_i p(x|\mu_i, \sigma_i), \quad \pi_i \geq 0, \quad \sum_i \pi_i = 1$$



Discriminative/Conditional model: Given a data set consisting of a predictor variable x and a response variable y, model the conditional distributions $p(y|x)$. Most classification and regression models are of this type.

- Example: Simple linear regression

$$y = \beta_0 + \beta_1 x + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

- Example: k-nearest neighbor classifier

Supervised and Unsupervised Learning

Supervised learning: Data is *labelled*.

- Discriminative models

Unsupervised learning: Data is *unlabelled*.

- Clustering
- Mixture models

Distinction is nebulous in situations where there is no clear notion of label, e.g., density estimation.