ETH zürich

Introduction to Machine Learning

Big Data for Public Policy

Christoph Goessmann
Law, Economics, and Data Science Group (Prof. Elliott Ash)

Course: 860-0033-00L Data Science for Public Policy: From Econometrics to AI, Spring 2024

28 March 2024

Student Projects

Paper or web app, 2-4 students per project

Timeline (see syllabus):

- **Topic (0%), due 9 April**. Please discuss with the instructors to select and confirm a topic, dataset, and approach.
- **Outline (5%), due 30 April**. 1/2 page outline of the motivation, related literature, data, and approach.
- **Presentation (20%), On 30 May**. Students will give a short presentation about their project toward the end of the course. It should include some preliminary analysis. 5 minutes, +2 minutes for each additional student in the group.
- **Deliverable (60%), due 21 July**. Paper OR web app.
  - A **paper** reporting on the project's analysis and results (Intro, Lit Review, Data, Methods, Results, and Conclusion). 2500+ words, +500 words for each additional student in the group.
  - **Web app**. The web app should visualize data, topics, model predictions… Points for creativity. The web app should be accompanied by a 1-page document explaining why this web app is of interest and the main challenges.
  - **Replication package (15%)**.

Repetition

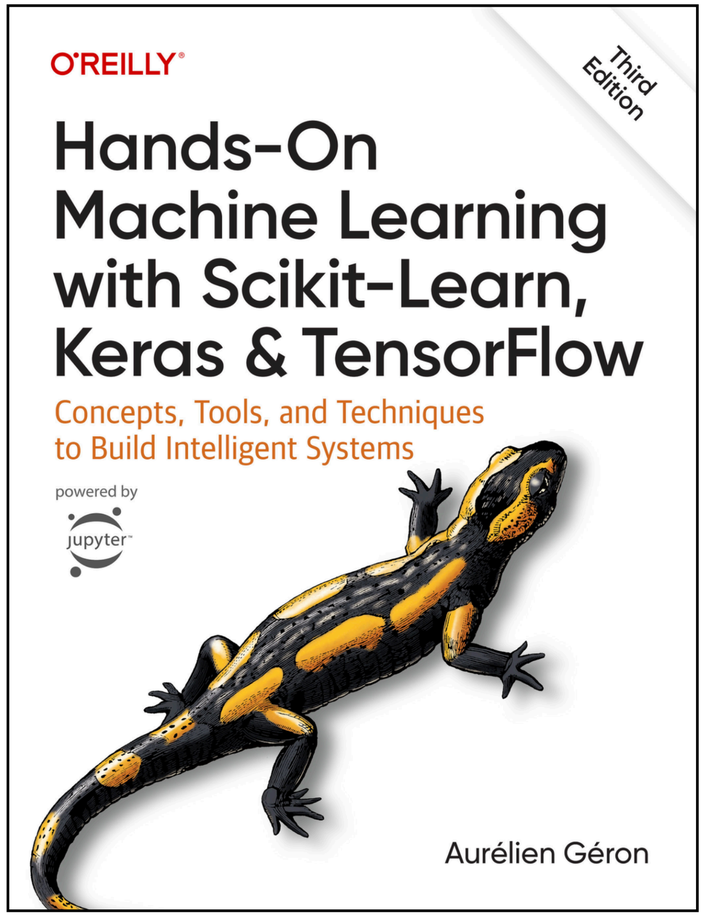
What did we do last time?

- Good practices for **data formats** / storage (e.g., *.csv, *.json)
- **APIs** / scraping
  - HTTP GET and HTTP POST
  - Proxies vs. VPN
  - Parsing
- Hands-on with **United Nations Statistics Division SDG API**
  - **Multi-threaded** API requests (to mitigate I/O bound → 12x faster)
  - **OLS regression** with API data (female representation in parliaments)

Tentative Outline for Today

- **What is machine learning?**
  - Machine learning vs. econometrics
  - Machine learning data classifications
- Performance metrics
- Loss functions / regularization

Literature Recommendation



Hands-On Machine Learning with Scikit-Lean, Keras & Tensorflow, Aurélien Géron, O'Reilly, Third Edition

- ISBN 978-1-098-12597-4
- Access at O'Reilly: click here and sign in with your ETH Zurich credentials
- Mobile app: Apple App Store, Google Play

What is Machine Learning?

Is a regression (e.g., the one that we did on female representation in parliaments two weeks ago) **machine learning?**

Yes.

What is Machine Learning?

A computer program [model] is said to **learn** from **experience E** with respect to some **task T** and some **performance [metric] P**, if its performance on T, as measured by P, improves with experience E.

Tom Mitchell, 1997

Algorithm vs. Data

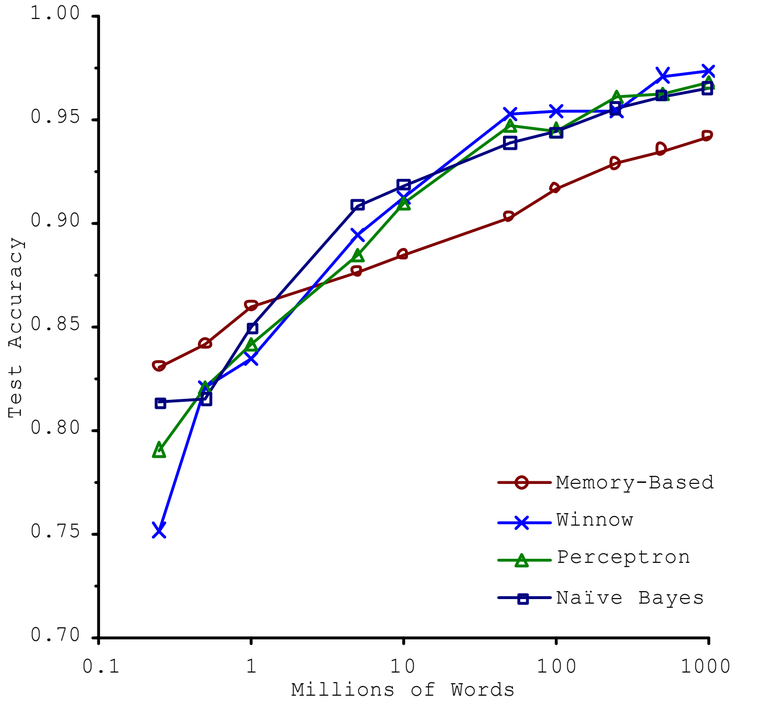


Figure 1. Learning Curves for Confusion Set Disambiguation

"We propose that a logical next step for the research community would be to direct efforts towards **increasing the size of annotated training collections**, while **deemphasizing the focus on comparing different learning techniques** trained only on small training corpora."

Source: Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL '01). Association for Computational Linguistics, USA, 26–33. https://doi.org/10.3115/1073012.1073017

Machine Learning vs Econometrics

Terminology and objectives

	Econometrics	Machine Learning
y	**dependent variable** regressand, outcome	**LABEL**
x	**independent variable(s)** regressor(s), explanatory variable(s)	**FEATURE(S)**
Objective	mostly **CAUSAL inference** → find and quantify causal relationships	**INFERENCE** → predict the label based on features

Task T

Some Examples

- Filter spam from ham
- Assign topics to news articles
- Detect tumors in brain scans
- Predict housing prices
- Detect fraud/outliers
- Dimensionality reduction
- Detect faces/people in pictures
- Transcribe audio to text
- Recommend similar products, songs, etc.

Experience E = Data

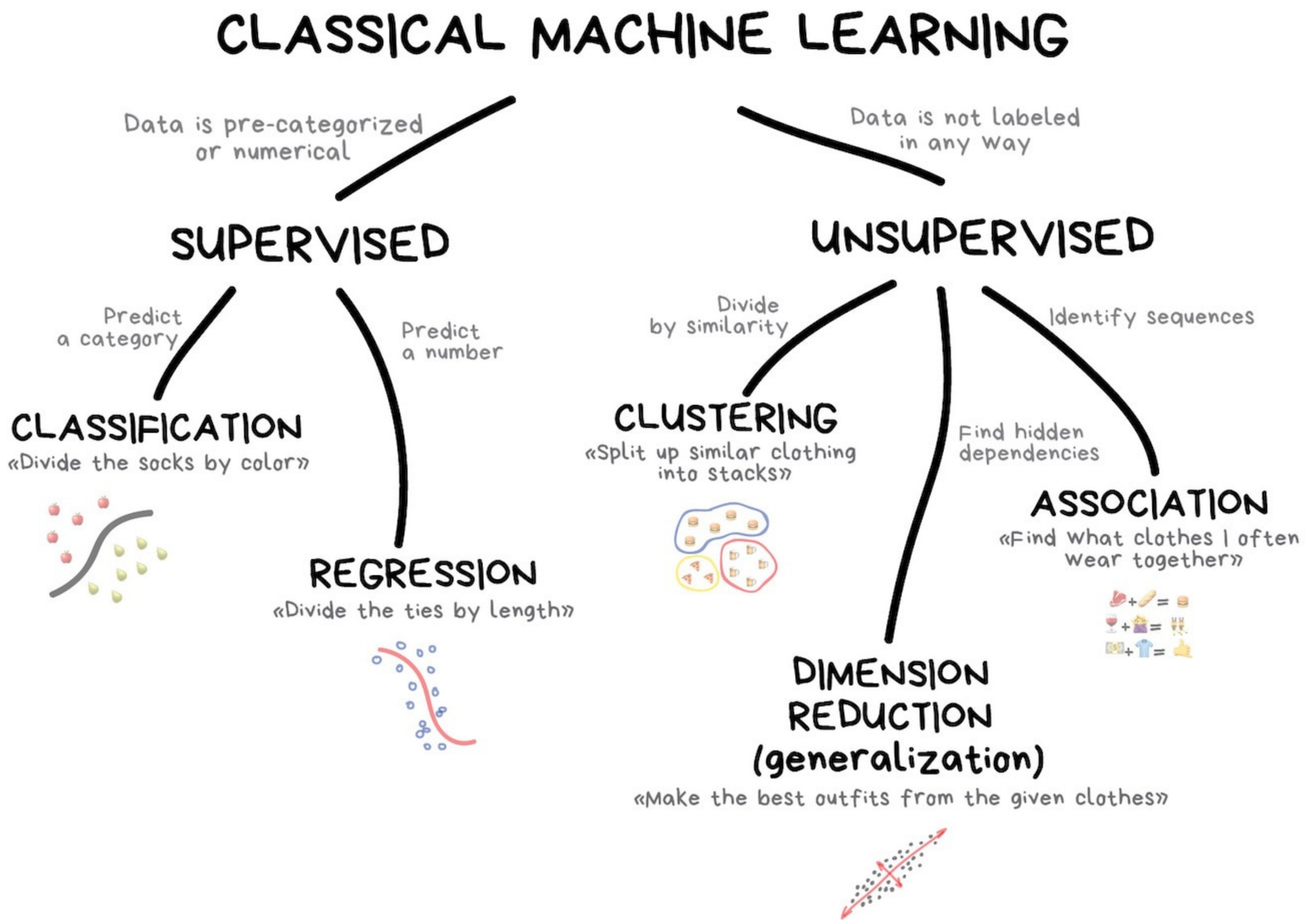
Common data classifications

Labeled Data

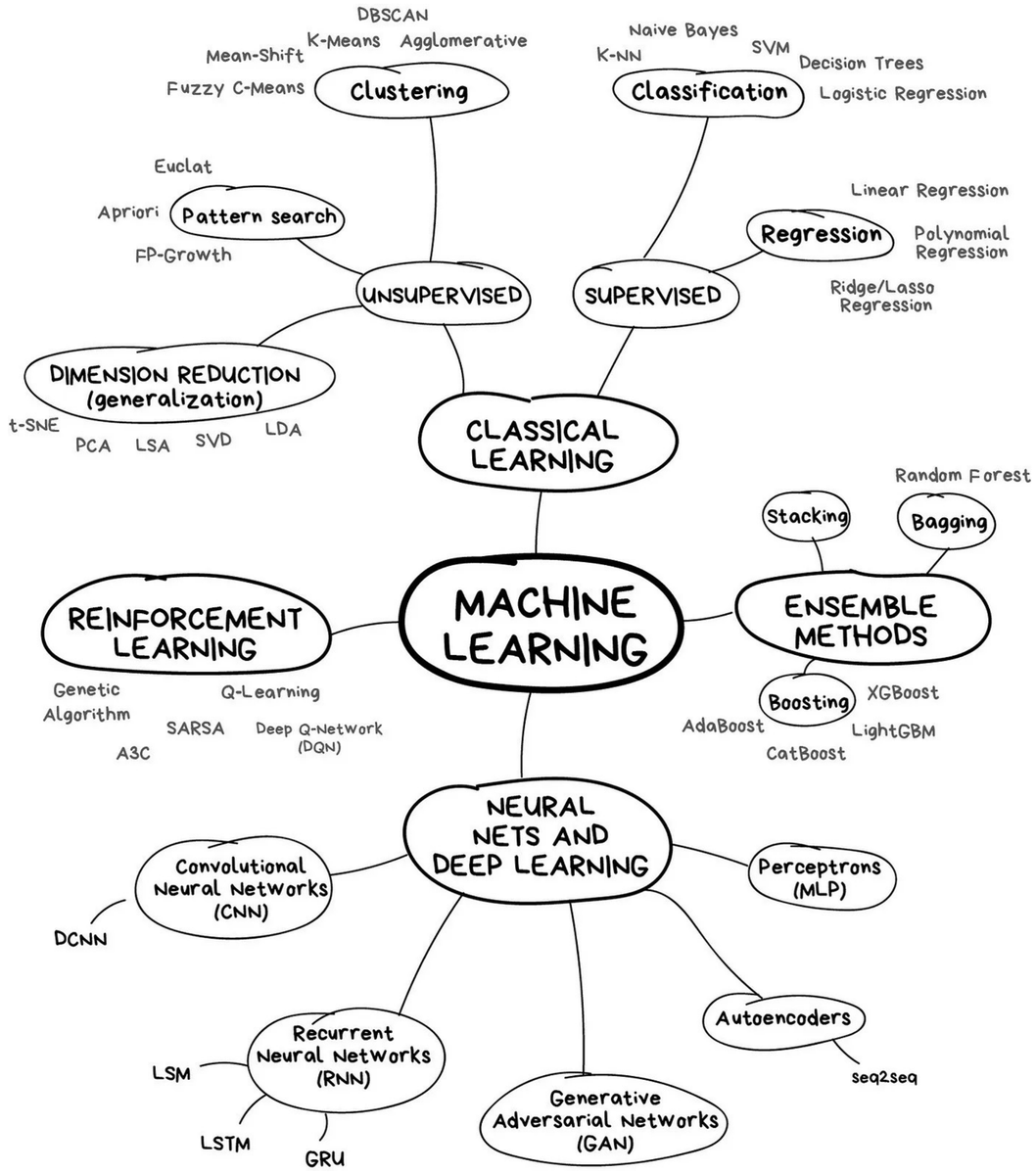
- Labels y (what we what to predict after having trained the model) are contained in the training data.
- If every label (e.g., cats, dogs, etc. for animal pictures) is contained equally often → **balanced** data, otherwise: **imbalanced.**
- **Supervised** machine learning requires labeled data.
  - **Continuous** vs. **discrete** labels:
    - **Regression** (linear and non-linear) for continuous labels
    - **Classification** for discrete labels

Unlabeled Data

- Labels y are not contained in the training data.
- **Unsupervised** machine learning.



Source: vas3k, https://vas3k.com/blog/machine_learning/index.html, accessed 30 March 2023



Source: vas3k, https://vas3k.com/blog/machine_learning/index.html, accessed 30 March 2023

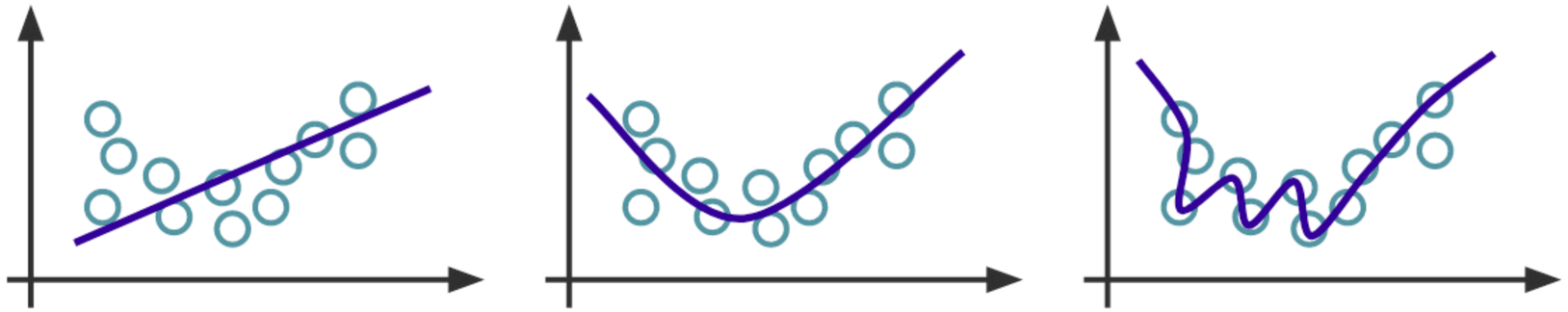
Performance Metrics P

How good is your model?

- We would normally **evaluate** a trained model's performance **on out-of-sample data** to see how well it generalizes.
- Typically, we split data into:
  - 80% **training data** and
  - 20% **test data** that the model has never see during training.
- Underfitting/overfitting would imply bad generalization/performance.

- Deployed ML models need to be evaluated regularly on **current** (representative of the data the model currently is fed with) test data.
- **Examples** for performance metrics:
  - **Regression**: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R2, Adjusted R2, ...
  - **Classification**: Accuracy, precision, recall, F1 score, ROC, ROC AUC, ...

Performance Metrics P



- Is this **classification** or **regression**?
- Which model **generalizes well**, is **overfitted**, or is **underfitted**?

Source: Leomaurodesenv, https://commons.wikimedia.org/wiki/File:Underfitting_e_overfitting.png, accessed 30 March 2023

Performance Metrics P

For regression

Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}(x_i, \theta) - y_i)^2$$

Where:

- n: number of observations in the **test data**
- $\hat{y}(x_i, \theta)$: label prediction for feature(s) x_i and model parameters θ
- y_i: true value for label

Slightly more interpretable:

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}(x_i, \theta) - y_i)^2}$$

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}(x_i, \theta) - y_i|$$

…

Performance Metrics P

For classification

Confusion Matrix		Predicted Values	
		Positive	**Negative**
Actual Values	**Positive**	**# True Positive (TP)**	**# False Negative (FN)**
	Negative	**# False Positive (FP)**	**# True Negative (TN)**

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + TP}$$

"Out of all predictions that I make, what proportion is correct?"

$$\text{Precision (for positive class)} = \frac{TP}{TP + FP}$$

"What proportion of the predicted positives are actual positives?"

$$\text{Recall (for positive class)} = \frac{TP}{TP + FN}$$

"What proportion of all actual positives do I predict correctly?"

Loss Functions

- Do not confuse the performance measure P with cost/error/**loss function** (smaller is better) or **utility function** (bigger is better)!
- **Loss/utility function** is used for **model optimization** during training.
- **Performance metric** is used after training to **evaluate the trained model's performance** on separate test data.
- The mathematical concept that you use (e.g., MSE, MAE) can be the same for loss/utility function and the performance metric, but they do not have to match.
- **OLS** tends to **over-fit to training data** and cannot handle **multi-collinearity**.
  - Introduce **regularization**!

Loss Functions

Regularization

Regularization introduces a **penalty** in the loss function or **reduces the degrees of freedom** for a model.

Ridge (ℓ^2) regression:

$$J = \|\hat{y}(X, \theta) - y\|_2^2 + \lambda \|\theta_j\|_2^2$$

Lasso (ℓ^1) regression:

$$J = \|\hat{y}(X, \theta) - y\|_2^2 + \lambda \|\theta_j\|_1$$

Elastic Net (Lasso+Ridge):

$$J = \|\hat{y}(X, \theta) - y\|_2^2 + \lambda_1 \|\theta_j\|_1 + \lambda_1 \|\theta_j\|_2^2$$

Where:

- $\hat{y}(x_i, \theta)$: label prediction (vector) for training feature(s) X (matrix)
- θ: model parameters/regression coefficients (vector)
- y: true values for labels (vector) from training data

In order to use regularization we normally have to scale input features.

End-of-Lecture Survey

ETH Edu App

Web app

https://eduapp-app1.ethz.ch/

iOS

Android