DATA 3464: Fundamentals of Data Processing

Exploratory data analysis

Charlotte Curtis

January 13, 2026

This week's topics

- Basic machine learning models
- The importance of understanding your data
- Exploratory visualizations
- Splitting and sampling

Machine learning

- To appropriately process the data, we need to know *why* we are doing it and what assumptions we're making
- Modern machine learning toolkits (such as scikit-learn) are so easy to use, they're easy to use inappropriately
- Goal: just enough understanding to use basic models **responsibly**

A selection of common models

Supervised

- Linear/logistic regression
- Decision trees
- Support vector machines

Unsupervised

Source: Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python. O'Reilly Media, Incorporated. Available at MRU Library.

No free lunch

Model evaluation: regression

Model evaluation: Classification

- **True positive**: predicted positive, label was positive (TP) ✓
- **True negative**: predicted negative, label was negative (TN) ✓
- **False positive**: predicted positive, label was negative (FP) ✗ (type I)
- **False negative**: predicted negative, label was positive (FN) ✗ (type II)
- **Accuracy** is the fraction of correct predictions, given as:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision and recall

- **Precision**: Out of all the positive **predictions**, how many were correct?

$$\text{precision} = \frac{TP}{TP + FP}$$

- **Recall**: Out of all the positive **labels**, how many were correct?

$$\text{recall} = \frac{TP}{TP + FN}$$

- **Specificity**: Out of all the negative **labels**, how many were correct?

$$\text{specificity} = \frac{TN}{TN + FP}$$

Confusion matrix

	Predicted Positive	**Predicted Negative**
True Positive	TP	FN
True Negative	FP	TN

- The axes might be reversed, but a good predictor will have strong diagonals
- There's also the **F1 score**, or harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

Check out the Wikipedia page for more ways of describing the same information

ROC Curves

- The receiver operating characteristic curve is a plot of the **true positive rate** (recall or sensitivity) vs. **false positive rate** (1 - specificity) as the detection threshold changes
- The diagonal is the same as random guessing
- A perfect classifier would hug the top left corner

> *Fun fact: the name comes from WWII radar operators, where true positives were airplanes and false positives were noise*