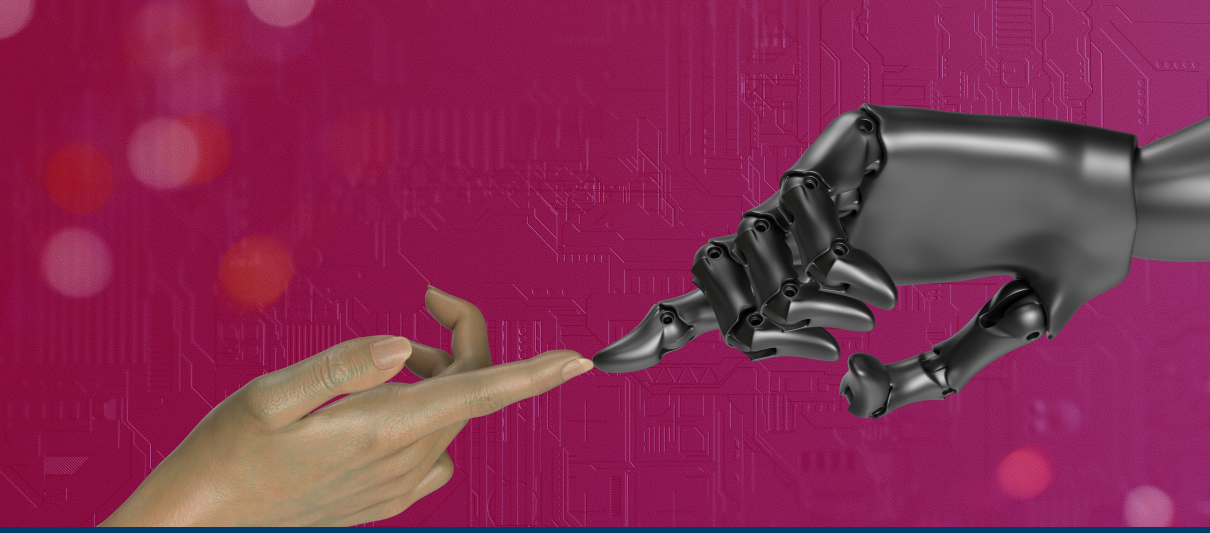


Hands-On Machine Learning

Dr. Derek Bridge | 2026/2027

UCC Coláiste na hOllscoile Corcaigh
University College Cork, Ireland

School of Computer Science and Information Technology
Scoil na Ríomheolaíochta agus na Teicneolaíochta Faisnéise

2. Supervised Learning

- We introduce some of the topics but using **synthetic data** — data that has been generated by a program that I wrote. For added realism the program injects some random noise into the data.
- The original idea for this data-generation program and its use in the later parts of the lecture come from Jake VanderPlas.

Prediction

- A lot of the time in Machine Learning, we want to create programs that make **predictions**.
- These programs are called **predictors** or **estimators**.
- (In everyday use, prediction is about the future; but we use the word more generally in Machine Learning.)

Regression

- **Regressors** predict continuous *numeric values*.
- E.g. given a description of a house, predict the selling price of the house.
- (Compare with **classification**.)

Dataset

A **dataset** is a set of **examples**.

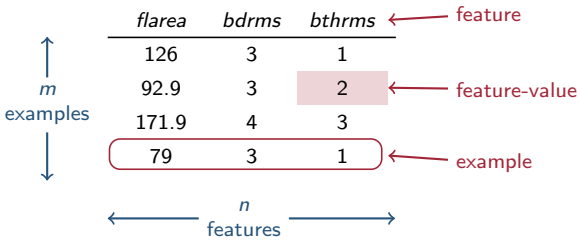
Examples

- An **example** is a single observation, representing an entity, event or situation.
- (An **example** may also be called an **instance** or a **sample**.)

Structured Dataset

- Each example fits a predefined schema.
- Most common is rows and columns, hence also called **tabular datasets**.
- (Compare with **unstructured dataset** and **multimodal dataset**.)

Tabular Dataset



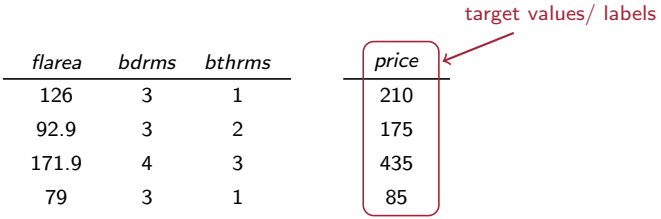
Labeled Dataset

- To *learn* how to make predictions, we need a **labeled dataset**.
- A labeled dataset is a set of **labeled examples**.

Labeled Example

- A **labeled example** is an example accompanied by its corresponding **label** or **target value**, i.e. the *actual* value we are trying to predict.
- E.g. a description of a house, but also its actual selling price.

Example Labeled Dataset



Notation

- We will refer to the two parts of this labeled dataset as $\boldsymbol{X}$ (uppercase, bold) and $\boldsymbol{y}$ (lowercase, bold).
- We will refer to an example (row) as $\boldsymbol{x}$ (lowercase, bold) and its corresponding target value/ label as y (lowercase, not bold).

Supervised Learning

- Given a labeled dataset;
- Learn to make predictions;
- Guided by the labels.

Model

In Supervised Learning, a **model** is a function (a formula, a set of rules, a procedure,...) that expresses the relationship between an object's features and the thing that is being predicted (the target value/label).

Training vs. Inference

- **Training**: Generalize from the training examples — learn a function (model) that captures the relationship between features and targets.
- **Inference**: Apply the learned function (model) to unseen examples — to make predictions.
- Training is also called **fitting a model** and inference is also called **prediction**.
- In `scikit-learn`, we have methods `fit` and `predict`.

Error Estimation

- We've learnt a model from a dataset.
- We want to know how well it will do in practice, once we start to use it for inference — to make predictions.
- This is called **error estimation**.
- We need one or more **evaluation metrics**.

Evaluation Metrics for Regression

Notation

- $\boldsymbol{X}$ is a dataset of examples.
- $\boldsymbol{x}$ is an example, $\boldsymbol{x} \in \boldsymbol{X}$.
- y is the *actual* target value for $\boldsymbol{x}$.
- $\hat{y}$ is the *predicted* target value for $\boldsymbol{x}$.
- $\overline{\boldsymbol{y}}$ is the mean of the actual target values.
- For regression, we measure the **error** (difference) between them: $\hat{y} - y$
- But we need to take the absolute value or square the error. Why?
- Why do some people prefer to square?
- We do this for each example in $\boldsymbol{X}$ and take the mean.

Mean absolute error (MAE)

$$MAE = \frac{1}{m} \sum_{\boldsymbol{x} \in \boldsymbol{X}, y \in \boldsymbol{y}} \text{abs}(\hat{y} - y)$$

Mean squared error (MSE)

$$MSE = \frac{1}{m} \sum_{\boldsymbol{x} \in \boldsymbol{X}, y \in \boldsymbol{y}} (\hat{y} - y)^2$$

Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{m} \sum_{\boldsymbol{x} \in \boldsymbol{X}, y \in \boldsymbol{y}} (\hat{y} - y)^2}$$

Coefficient of determination (R^2-score)

$$R^2\text{-}score = 1 - \frac{\sum_{\boldsymbol{x} \in \boldsymbol{X}, y \in \boldsymbol{y}} (\hat{y} - y)^2}{\sum_{\boldsymbol{x} \in \boldsymbol{X}, y \in \boldsymbol{y}} (\overline{\boldsymbol{y}} - y)^2}$$

Holdout

- **Holdout** is a method for *partitioning* a dataset:
  - Shuffle the dataset — to ensure a random split.
  - Partition the dataset into two:
    - **training set** (e.g. the first 80% of the full dataset);
    - **test set** (the rest of the full dataset).
  - Train the model on the training set.
  - Test the model on the test set.
- This is called **holdout**, because the test set is withheld (held-out) during training.
- We need the training and test sets to be representative of the full dataset — otherwise, informally speaking, you might get 'lucky' or 'unlucky'.
- Holdout ideally needs a large dataset. With a small dataset, there is a greater chance that they will be unrepresentative.

Training Error and Test Error

- We learn a model by training on the training set (of course).
- **Training error**: This is where we evaluate also on the training set.
- **Test error**: This is where we evaluate on the test set.
- Generally, we are more interested in the test error — this shows how the model performs on **unseen data** and therefore how it is likely to perform when we deploy it.

Leakage

- It is essential that the test set is not used in any way to create the model. *Don't even look at it!*
- 'Cheating' is called **leakage**.
- Leakage will mean that you will probably underestimate the error of your model when it faces unseen data.
- The most obvious example of leakage is using the same data for training and testing.
- But watch out for more subtle examples too!

Underfitting and Overfitting

- **Underfitting**: the model is not complex enough.
- **Overfitting**: the model is too complex.