# Supervised Machine Learning: Foundations

Arnav Gupta

October 18, 2024

## Contents

## 1 Learning

The ability to improve behaviour based on experience. Improvements can be:

- **range** of behaviours is expanded: agent can do more
- **accuracy** on tasks is improved: agent can do things better
- **speed** is improved: agent can do things faster

Components of a learning problem:

- **task**: behaviour being improved
- **data**: experiences used to improve performance in the task
- **measure of improvement**

Common learning tasks include:

- **supervised classification**: given a set of pre-classified data, classify a new instance
- **unsupervised learning**: find natural classes for examples
- **reinforcement learning**: determine what to do based on rewards and punishments
- **transfer learning**: learn from an expert
- **active learning**: learner actively seeks to learn
- **inductive logic programming**: build richer models in terms of logic programs

Learning tasks can be characterized by the feedback given to the learner:

- for supervised learning, what has to be learned is specified per example
- for unsupervised learning, no classifications given so learner must discover categories and regularities in data
- for reinforcement learning, feedback occurs after a sequence of actions

## 1.1 P/N Agents

Measure of success is how well the agent performs on new, unseen examples.

For two agents solving a binary classification task:

- $P$ claims the negative examples seen are the only negative examples (all others positive)
- $N$ claims the positive examples seen are the only positive examples (all others negative)

Both agents correctly classify every training example, but disagree on every other example. They use training data as their model, needing an exact match.

### 1.1.1 Bias

Tendency to prefer one hypothesis over another. Necessary to make predictions on unseen data.

To make predictions on unseen data, bias is necessary. Good biases must be tested to find out their quality.

### 1.2 Learning as Search

Given a representation and bias, learning is search through all possible representations looking for the representation(s) that best fit the data, given the bias.

Learning algorithm is made of:

- search space (possible representations)
- evaluation function
- search method

## 2 Supervised Learning

Supervised learning requires:

- a set of input features $X_1, \ldots, X_n$
- a set of target features $Y_1, \ldots, Y_k$
- a set of training examples with values given for input features and target features
- a set of test examples where only values for input features are given

Goal: Predict values of target features for test examples using

- classification when target features are discrete
- regression when target features are continuous

### 2.1 Noise

Data can have:

- some features assigned the wrong value
- inadequate features to predict classification
- missing features

**Overfitting**: distinction appears in training data but not in test data, due to random correlations

## 3 Measures of Error

For a feature $Y$ and example $e$:

- $Y(e)$ is the value of feature $Y$ for $e$
- $\hat{Y}(e)$ is the predicted value of feature $Y$ for $e$
- **error** of prediction: measure of how how close $Y(e)$ and $\hat{Y}(e)$ are

Let $E$ be a set of examples and $T$ be a set of target features:

- **absolute error**

$$\sum_{e \in E} \sum_{Y \in T} \left| Y(e) - \hat{Y}(e) \right|$$

- **sum of squares error**

$$\sum_{e \in E} \sum_{Y \in T} \left( Y(e) - \hat{Y}(e) \right)^2$$

- **worst case error**

$$\max_{e \in E} \max_{Y \in T} \left| Y(e) - \hat{Y}(e) \right|$$

- **cost-based error**: takes into account costs of various errors

For the case where target features are $Y(e) \in \{0, 1\}$ and predicted features are $\hat{Y}(e) \in [0, 1]$:

- **likelihood of the data**:

$$\prod_{e \in E} \prod_{Y \in T} P(\hat{Y}(e) \mid Y(e))$$

$$\prod_{e \in E} \prod_{Y \in T} \hat{Y}(e)^{Y(e)} (1 - \hat{Y}(e))^{(1 - Y(e))}$$

- **entropy** or **negative log likelihood**

$$-\sum_{e \in E} \sum_{Y \in T} \left[ Y(e) \log(\hat{Y}(e)) + (1 - Y(e)) \log(1 - \hat{Y}(e)) \right]$$

## 3.1 Precision and Recall

**Recall** or **sensitivity**

$$\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

**Specificity**

$$\frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$

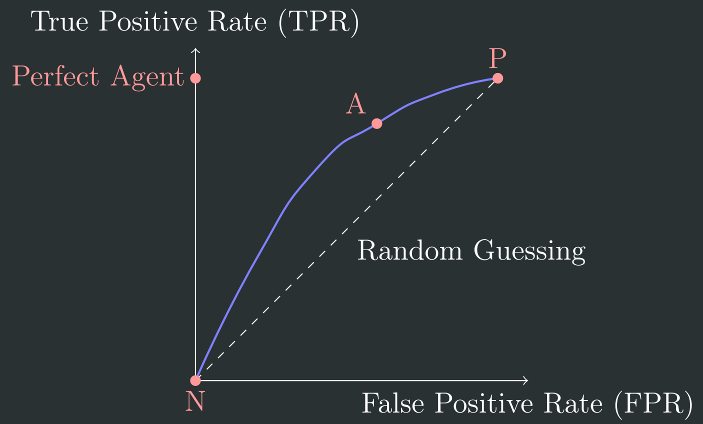
**Precision**

$$\frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

**F1-measure**

$$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The **receiver operating curve** gives the full range of performance of an algorithm across different biases.


True Positive Rate (TPR)
Perfect Agent
P
A
Random Guessing
N
False Positive Rate (FPR)