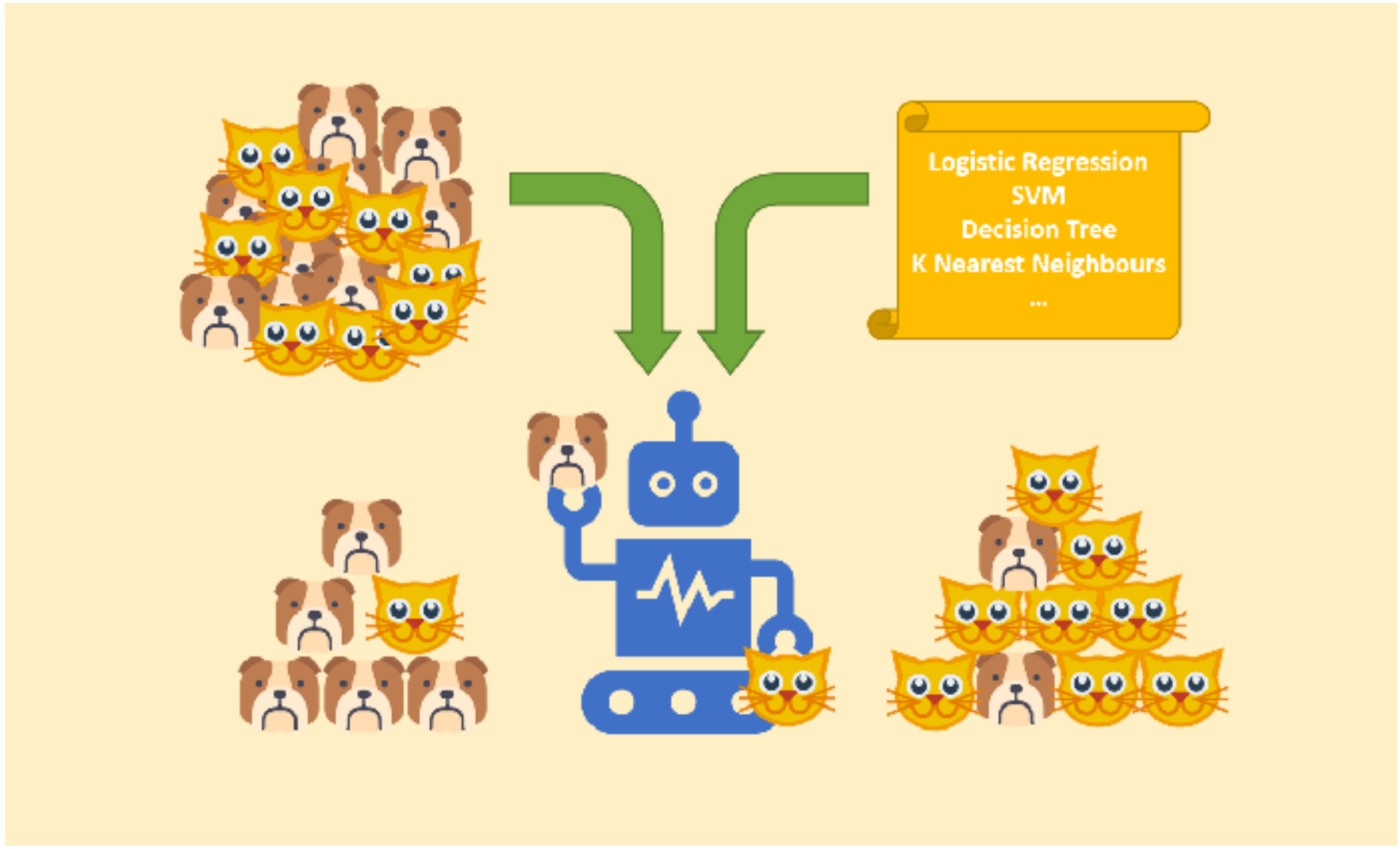


Introduction to Classification

Leandro L. Minku

Supervised Learning of Classification Problems

- Supervised learning:
  - Given a set of training examples

$$\mathcal{T} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \cdots, (\mathbf{x}^{(N)}, y^{(N)})\}$$

where $(\mathbf{x}^{(\mathbf{i})}, y^{(i)}) \in \mathcal{X} \times \mathcal{Y}$ are drawn from a fixed albeit unknown joint probability distribution $p(\mathbf{x}, y)$.

  - Learn a (predictive) model $f : \mathcal{X} \to \mathcal{Y}$ able to generalise to unseen (test) examples of the same probability distribution $p(\mathbf{x}, y)$.

- Regression problem: $\mathcal{Y}$ is $\mathbb{R}$.
- Classification problem: $\mathcal{Y}$ is a set of categories / classes.
  - When $\mathcal{Y}$ is a set of size 2, we call the problem as a binary classification problem.

Independent variables

Dependent variable

What About $\mathcal{X}$?

d-dimensional space, where each dimension can be:

- Numeric:
  - E.g., age, salary.
- Ordinal:
  - E.g., expertise in {low, medium, high}.
- Categorical:
  - E.g., car in {fiat, volkswagen, toyota}.

Different dimensions may be of different types.

Examples of Classification Problems

Credit Card Approval:

- Prediction of whether a customer will pay their credit card bills or default their payments.
- Based on independent variables such as age, gender, salary, type of bank account, etc.
- Predictive models can be created based on data describing previous customers.

Supervised Learning Problem

- E.g.: credit card approval

$$\mathcal{T} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \cdots, (\mathbf{x}^{(N)}, y^{(N)})\}$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \begin{matrix} \longrightarrow \text{age} \\ \longrightarrow \text{salary} \\ \\ \longrightarrow \text{gender} \end{matrix} \qquad \mathbf{x}^{(1)} = \begin{pmatrix} 21 \\ 1500 \\ \vdots \\ \text{male} \end{pmatrix}$$

$$\mathbf{x}^{\mathrm{T}} = (x_1, x_2, \cdots, x_d)$$

$$\mathbf{x} = (x_1, x_2, \cdots, x_d)^{\mathrm{T}}$$

$$y^{(1)} = \text{good}$$

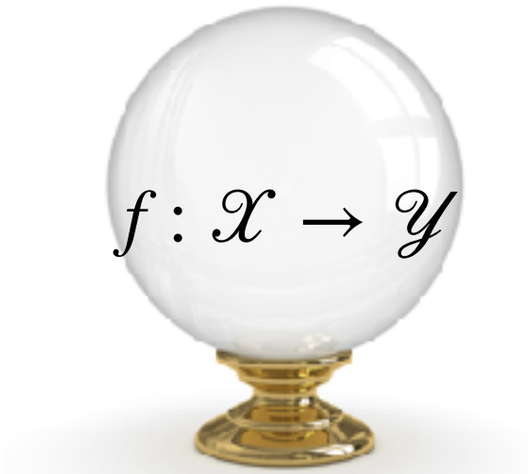
$y \longrightarrow$ good/bad

Supervised Learning

Training Data / Examples

$\mathcal{T} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \cdots, (\mathbf{x}^{(N)}, y^{(N)})\}$

→ Machine Learning Algorithm →



Predictive Model

New example **a** to be predicted →

→ Prediction

$\hat{y} = f(\mathbf{x})$

Predictions will not always be correct.

Examples of Classification Problems

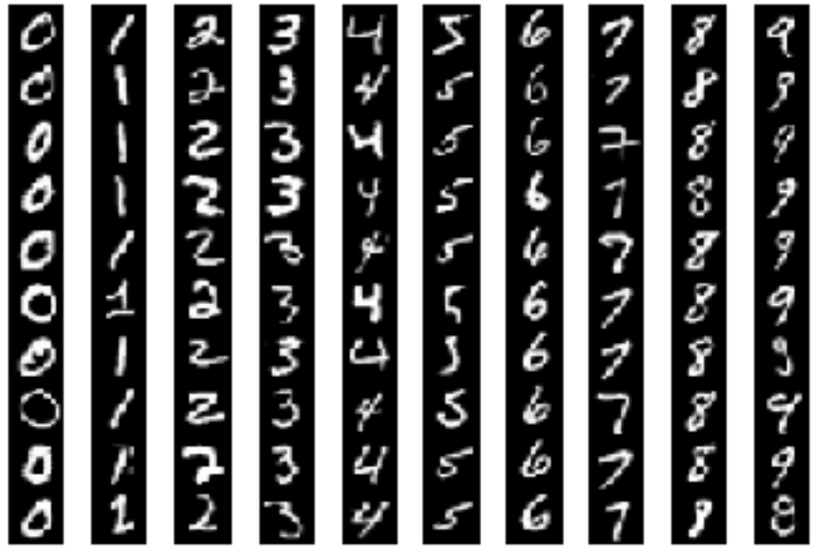
Breast Cancer Prediction:

- Predict whether a person does or does not have breast cancer.
- Based on independent variables such clump thickness, uniformity of cell size, cell thickness, etc.
- Predictive models can be created based on data describing previous patients.

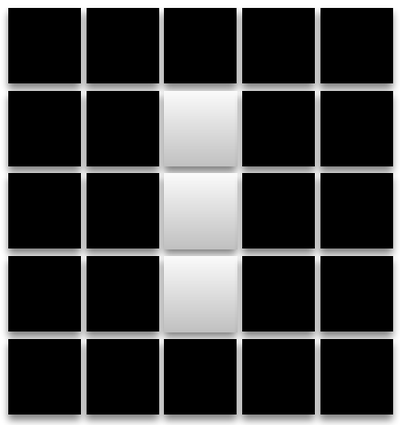
Examples of Classification Problems

Handwritten digits recognition:

- Predict which digit is written.
- Based independent variables representing the colour of the pixels composing the image of the handwritten digit.
- Predictive models can be created based on previous handwritten digits and their labels.



Vectorial Format of Independent Variables



An image is a matrix of pixels,
but we can convert it to a vector $\mathbf{x}^{\mathrm{T}}$.

Equivalent Terms

- Independent variable, input attribute, input variable.
- Dependent variable, output attribute, output variable, label (for classification).
- Predictive model, classifier (for classification).
- Learning a model, train a model.
- Training examples, training data.
- Example, observation, data point, instance (more frequently used for test examples).

Quiz

- In a classification problem…
  - the independent variables are categorical.
  - the dependent variable is categorical.
  - the independent variables are numeric.
  - the dependent variable is numeric.
  - both the independent and dependent variables are categorical.
  - both the independent and dependent variables are numeric.

Further Reading

- Essential:
  - Iain Styles' notes on "Classification and k-Nearest Neighbours", page 2.