Chapter 1

- **Machine Learning -** automated methods for automatically detecting patterns in data
  - Perform decision making or predict future data
- **Long tail** data **-** a pattern in data where few things are very common, but most are rare
- $p(y|\mathbf{x})$ **Supervised Learning (Predictive) -** learn a mapping from inputs **x** to outputs **y** given a **training set**
  - **Classification -** predicting a **categorical variable** (boundary)
  - **Regression -** predicting a real value (line of best fit)
    - **Ordinal Regression -** predicting grade A - F
- $p(\mathbf{x})$, **Unsupervised learning -** finding patterns in data **without labels**
  - **Knowledge Discovery**, less well-defined problem
- **Reinforcement Learning -** learning how to act or behave given an occasional **reward/punishment**
- **Generalisation -** making correct predictions on novel inputs
- **Probabilistic Prediction -** for ambiguous cases, it is desirable to return a probability
- **Maximum a Posteriori -** Predicting the **mode**, the class with the most probable label
  - $$\hat{y} = \hat{f}(\mathbf{x}) = \operatorname*{argmax}_{c=1}^{C} p(y = c|\mathbf{x}, \mathcal{D})$$
- **p^ -** our estimated output class
  - $p(\hat{y}|\mathbf{x}, \mathcal{D})$ **-** If our predicted class has a low probability, it might be better to say IDK
- Useful for **medicine**, **finance**, or where **risk** quantifying is important
- **Click Through Rate CTR -** likelihood a user would click on an ad, usually based on search history

1.2 Supervised Learning

- **Conditional Density Estimation -** $p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$
  - Probability of class y **conditional on input x** and parameters **θ**

1.2.1.3

- **Bag Of Words -** define $\mathbf{x_{ij} = 1}$ if word **j** occurs in document **i**
- **Exploratory Data Analysis -** plotting data and looking for patterns in the features by visualising them
- **Invariant -** A model should be robust to slight changes in previously seen examples

1.3 Unsupervised Learning

- **Knowledge Discovery -** finding interesting structure in the data
- **Unconditional Density Estimation -** predict $p(\mathbf{x}_i|\boldsymbol{\theta})$
- **Multivariate probability models -** our **x** is a vector of features, so we need multivariate probability models, that will give us probability distributions for each feature

1.3.1 Discovering Clusters

- First step - finding **p(K|D)** - probability distribution over **number of clusters**
  - Finding how likely it is that there are **K** number of classes
  - We approximate **K** by estimating the **mode** $K^* = \arg\max_K p(K|\mathcal{D})$.
    - Supervised classes tell us this info before hand
- Second step - estimate which cluster is the most likely, $z_i^* = \operatorname{argmax}_k p(z_i = k|\mathbf{x}_i, \mathcal{D})$
  - $\mathbf{z_i}$ is the cluster that $\mathbf{x_i}$ has been assigned to
    - **z** is a hidden / **latent variable -** variables **inferred** from other variables

1.3.2 Latent Factors

Latent Factors - a small number of degrees of variability (things that creates our data), even if our data is represented by more dimensions than the number of latent factors

- We can compress our data down to just the latent factors
- Underlying parameters that describe most of the variability

1.3.4 Matrix Completion / Imputation - plugging in the gaps in data matrices

- **Image Inpainting -** fixing gaps in images with realistic texture
- **Collaborative Filtering -** predicting which movies people will want to watch based on other people's tastes

1.4 Basic Concepts of Machine Learning

Parametric Model - A model that has a fixed number of parameters, regarding of the amount of training data

- Stronger inductive prior (makes stronger assumptions about data distributions)
- Faster to use

Non-Parametric Model - A model that grows in number of parameters with more and more training data

- More flexible
- Computationally intractable for large N
-