Module 4: Statistics for Machine Learning

Learning outcomes

1. Identify the real-world applications of statistics in machine learning.
2. Identify the Python functions used to calculate maximum likelihood estimation.
3. Calculate maximum likelihood estimation in Python based on existing data.
4. Identify outliers in a data set.
5. Analyse a data set to identify mean, median, standard deviation and outliers.
6. Analyse the impact of removing outliers from a data set.
7. Use the regression formula to calculate regression coefficient for a data set.
8. Use the correlation formula to calculate correlation coefficient for a data set.
9. Analyse the relationships among data in a data set.
10. Apply concepts of outliers and correlations to solve a business problem.
11. Apply the bootstrapping resampling technique in Python to determine the accuracy of statistical summaries.
12. Examine the foundational concepts of statistics in machine learning, specifically the accuracy of statistical estimations and inferences.
13. Evaluate the appropriate machine learning competitions for career advancement.

Key statistical methods for machine learning (ML)

1. Maximum likelihood estimation (MLE)

MLE is a fundamental statistical method used to estimate the parameters of a probability distribution that best explains the observed data. It maximises the likelihood function, ensuring the model is most probable given the data.

- **Likelihood function:**

$$L(\theta; X) = \prod_{i=1}^{n} f(x_i; \theta)$$

- **Log-likelihood function:**

$$l(\theta; X) = \log L(\theta; X)$$

- **Optimisation:**
  Find θ that maximises $L(\theta; X)$ or $l(\theta; X)$. Many optimisation algorithms minimise functions, so this can be reframed as minimising the negative log-likelihood $-l(\theta; X)$.

- **Common distributions using MLE include:**

Distribution	Probability density function	MLE estimate
Bernoulli distribution: estimates the probability of a binary outcome (success or failure)	$P(X = x; p) = p^x(1-p)^{1-x}$	$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$
Binomial distribution: estimates the probability of success in a fixed number of independent trials	$\mathrm{P(X = x; n, p)} = \binom{\mathrm{n}}{\mathrm{z}}\mathrm{p^x(1-p)^{n-x}}$	$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n.N} \| 1$
Poisson distribution: estimates the average rate of events occurring in a fixed interval of time or space	$P(X = x; \lambda) = \frac{\lambda^z exp(-\lambda)}{x!}$	$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i (sample\ mean)$
Normal distribution: estimates the average value and spread (mean and variance) of continuous data	$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i\ (sample\ mean)$ $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\ (sample\ variance)$
Gamma distribution: models waiting times or accumulated damage and other non-negative quantities	$f(x; \alpha, \beta) = \frac{x^{\alpha-1} exp(-x/\beta)}{\beta^\alpha T(\alpha)}$	The MLE estimators are obtained by solving the likelihood equation, often requiring numerical methods for non-integer a.

- **Central limit theorem:**
  States that the sum of many independent random variables tends to follow a normal distribution, regardless of the original distribution

2. **Outlier detection**
   Outliers are data points that significantly deviate from the rest of the data set. Detecting them is crucial in data analysis, as they can distort models or indicate anomalies worth investigating.
   - **Methods:**
     - **Z-score:** measures how far a data point is from the mean in terms of standard deviations
     - **Interquartile range (IQR):** identifies outliers based on the spread of the middle 50 per cent of the data ($IQR = Q3 - Q1$)
     - **Outliers are typically defined as:**
       - **Low outliers:** $x < Q1 - 1.5 \times IQR$
       - **High outliers:** $x > Q3 + 1.5 \times IQR$

3. **Regression analysis**
   Regression is a statistical method used to model relationships between variables. It helps in making predictions and understanding dependencies between features in data.
   - **Linear regression model:** models the relationship between an independent variable (X) and a dependent variable (Y)
   - **Least squares estimation:** determines the best-fitting line by minimising the sum of squared differences between observed and predicted values

4. **Correlation coefficients**
   Correlation measures the strength and direction of the relationship between two variables. A positive correlation means both variables increase together, while a negative correlation means one increases as the other decreases.
   - **Pearson correlation formula:**
     - Values range from **−1 (strong negative correlation)** to **+1 (strong positive correlation).**

5. **Bootstrapping**
   Bootstrapping is a resampling technique used to estimate the accuracy and variability of statistical measures. It is particularly useful when working with limited data.
   - **Process:**
     - Randomly sample data with replacement to create new data sets.
     - Compute statistical measures (e.g. mean or correlation) for each sample.
     - Repeat many times to estimate variability and confidence intervals.
   - **Applications:** used in confidence interval estimation, bias reduction and improving ML model reliability