



NLP and classification

Empowering insights through classification

5 WEEKS

175 HOURS

10 LESSONS

In this module, we will embark on a journey through the realms of **natural language processing (NLP)** and **classification**. We'll begin by delving into **text preprocessing techniques**, including tokenisation, stemming, lemmatisation, and **feature extraction methods** such as bag-of-words and n-grams. We will explore **various classification algorithms** through practical implementations, understand their operational mechanisms, and learn how to evaluate their performance effectively.

From **logistic regression and classification metrics** to **tree-based methods**, **support vector machines (SVM)**, and **neural networks**, we'll equip ourselves with a diverse toolkit for tackling real-world classification tasks. Using **model-tuning** examples and exercises, we'll gain valuable insights into the iterative process of building and refining classification models.

Module objectives

Natural language processing

Gain a fundamental understanding of **text preprocessing methods** such as **tokenisation**, removing **stopwords**, **stemming**, **lemmatisation**, and **removing punctuation**. Understand feature extraction techniques such as **bag-of-words** and **n-grams** to convert text data into numerical representations.

Classification techniques and methods

Learn how to **implement** and **evaluate various classification techniques** using sklearn, including **logistic regression**, **decision trees**, **random forests**, **SVM**, **naive Bayes**, and **k-nearest neighbours**. Gain insights into model operations, **visualise SVM hyperplanes**, comprehend the **log loss function**, and deploy **neural networks** using **TensorFlow** and **Keras** for optimal classification outcomes.

Model evaluation and metrics

Acquire a foundational understanding of classification techniques, including **performance evaluation**, **class imbalance handling**, and **multiclass challenges**. Apply **feature subset selection** methods and master **hyperparameter tuning** across various models. Develop proficiency in employing metrics such as **accuracy**, **precision**, **recall**, and **F1-score** for effective model evaluation.

Model selection

Learn how to **construct diverse classification models**, conduct **cross-validation** to assess **performance robustness**, and employ visualisation techniques to **interpret results** effectively. Gain skills in **selecting the most suitable model** for a dataset based on performance evaluation metrics and graphical representations, enhancing the ability to **make informed model choices** in practical applications.

Learning activities

Participating in various learning activities will **enhance our understanding** of NLP and classification, by fostering a diverse skill set for **real-world challenges** through **hands-on problem-solving** and **authentic projects**.

We learn by doing. We'll work on practical problem-solving and real-world projects.

Learn

Watch animated videos and explore practical examples to learn NLP and classification concepts.



12

Sketch videos



13

Examples

Apply

Practise NLP and classification modelling and assessment by following detailed step-by-step guides and applying these techniques to real-world scenarios.



5

Exercises

Assess

Test and track your understanding of NLP and classification and their application.



12

KQ assessments



5

Graded assessments



NLP and classification

Empowering insights through classification.

Week 1

Lesson: Natural language processing

In this lesson, we will explore text preprocessing methods such as **tokenisation**, **removing stopwords**, **stemming**, **lemmatisation**, and **removing punctuation**. We will also examine feature extraction techniques such as **bag-of-words** and **n-grams** to convert text data into numerical representations.

- ✓ Understand the importance of **text preprocessing** in natural language processing tasks.
- ✓ Learn how to **tokenise text data** into words or tokens.
- ✓ Implement techniques to **remove stopwords and punctuation from text data**.
- ✓ Apply stemming and lemmatisation to **extract the root forms of words**.
- ✓ Create a bag-of-words representation to **quantify the occurrence of words** in text data.
- ✓ Explore the concept of n-grams to **capture combinations of words** in text data.

Week 2

Lesson: Logistic regression

In this lesson, we'll learn how to **implement logistic regression models** with sklearn, tackle real-world classification problems, **evaluate model performance**, as well as explore the **limitations of logistic regression**.

- ✓ Distinguish between **binary classification and regression**, understanding when and why to use each.
- ✓ Grasp the concept of logistic regression and its effectiveness in **binary classification scenarios**.
- ✓ Implement a logistic regression model using **sklearn** and **evaluate its performance** on real-world data.
- ✓ Develop skills in **preprocessing data**, **fitting logistic regression models**, and applying them to **solve classification problems**.

Lesson: Classification metrics

This lesson introduces the fundamental **metrics** used to **assess classification models**, such as accuracy, precision, recall, and the F1 score. We will learn when to use them and the benefits gained in doing so.

- ✓ Understand how the issue of **class imbalance** affects model evaluation.
- ✓ Know how to **evaluate the performance of a classification model** by interpreting a confusion matrix and a classification report.
- ✓ Understand and implement various **binary classification metrics**.

Lesson: Model improvements

In this lesson, we will cover a range of techniques, from **feature subset selection** and **hyperparameter tuning**, to addressing **class imbalance** and introducing **multiclass classification** strategies. We'll refine logistic regression models, utilise **advanced tuning methods**, and effectively manage dataset peculiarities for improved model performance.

- ✓ Understand and apply **feature subset selection** techniques such as **variance threshold**, select **KBest**, and **forward/backward stepwise selection**.
- ✓ Master **hyperparameter tuning** for various models, including logistic regression, random forests, and neural networks.
- ✓ Rebuild and evaluate logistic regression models using **performance metrics** to enhance prediction accuracy.
- ✓ Address class imbalance using **threshold adjustments** and **resampling** methods.
- ✓ Navigate the complexities of **multiclass classification**.
- ✓ Explore various metrics and concepts used for **evaluating classification models**.

Week 3

Lesson: Tree-based classification methods

In this lesson, we **refresh our understanding** of how decision trees and random forests work. We also look at their relevance and application in **classification contexts** and how we can implement them using sklearn to solve classification problems.

- ✓ Recall the terms and the process used for **training decision trees**.
- ✓ **Compare** the decision tree **training process** in regression and classification settings.
- ✓ Build and evaluate **tree-based models for classification**.

Lesson: Support vector classification

In this lesson, we will delve into **support vector machines (SVM)** and the art of **model tuning**. We will gain insights into the working principles of SVMs and the role of kernels. By exploring both **linear** and **non-linear SVM** classifiers using sklearn, we will explore the **significance of hyperplanes** and their **visualisation**. Moreover, we will learn to **fine-tune SVM models** effectively using **GridSearchCV**, thereby **optimising** model performance for real-world applications.

- ✓ Understand the **fundamentals of support vector machine (SVM)** models and their operational mechanisms.
- ✓ Implement SVM classifier models with **linear and radial basis function kernels** using the sklearn library.
- ✓ **Visualise hyperplanes** in SVM classifier models to understand their significance in **decision boundary determination**.
- ✓ Learn to **tune SVM models** using the GridSearchCV function, optimising model parameters for improved performance.

Lesson: Nearest neighbours and naive Bayes

In this lesson, we will explore two fundamental algorithms used in classification: **K-nearest neighbours (KNN)** and **naive Bayes classifiers**. By understanding the inner workings of these models and their practical implementations, we will gain valuable insights into how to **effectively classify data**.

- ✓ Explain the concept of **overfitting** and the **importance of regularisation** in machine learning.
- ✓ Implement **data scaling techniques** to improve model performance.
- ✓ Apply **ridge and LASSO regression methods** to prevent overfitting.

Week 4

Lesson: Hyperparameter tuning and model validation

In this lesson, we will delve into the intricacies of **hyperparameters and model validation**. We'll explore the significance of tuning hyperparameters and learn how to find optimal settings using a **grid search**. By understanding these concepts, we'll enhance our ability to **fine-tune models for optimal performance**.

- ✓ Understand the concept of **hyperparameter tuning** and its importance in **model optimisation**.
- ✓ Utilise **grid search** methodology to **find optimal hyperparameters for models** such as KNN and SVMs.
- ✓ Apply **model validation techniques** to assess and optimise the performance of machine learning models.

Lesson: Neural network classifiers

In this lesson, we will explore the fundamentals of **artificial neural network (ANN)** classifiers. We will introduce the architecture and components of ANN classifiers, including **layers, neurons, weights, and activation** functions. We will use **TensorFlow** and **Keras** to construct and train neural network models, focusing on classification tasks. Additionally, we will delve into model evaluation techniques, including the use of validation data to guide model training effectively.

- ✓ Obtain a basic understanding of the architecture and fundamental elements of an **artificial neural network**.
- ✓ Understand how to use **TensorFlow layers** to build a neural network architecture.
- ✓ Understand how a model is trained and evaluated using a **validation split**.
- ✓ Implement an effective neural network for classification using **Keras**.

Lesson: Classifier model selection

In this lesson, we'll construct several classification models, conduct **cross-validation**, and **visualise** the outcomes. The focus will be on determining the most suitable model for a dataset based on **performance metrics**. This iterative approach exemplifies the machine learning principle of rapid **experimentation** to refine model solutions effectively. Through hands-on exercises, we will gain valuable insights into **model selection and validation processes**.

- ✓ Build multiple types of classification models, including **logistic regression, k-nearest neighbours (KNN), support vector machines (SVM), decision trees** and **AdaBoost**.
- ✓ Perform **cross-validation** to assess the robustness of the models.
- ✓ Visualise the **results** of different classifiers.

Week 5

Exam: NLP and classification

It is time to review all the work that's been covered up to now and **test our knowledge**. Make sure to **cover each week's section in detail** before attempting this exam! It will be a combination of **theoretical** and **practical questions**, aimed at testing the **general understanding of concepts**, as well as the **application** and **interpretation of our new skills**.

Module summary


Throughout this module, we've delved into the intricacies of **natural language processing and classification**. We began by understanding the importance of **text preprocessing** and **feature extraction techniques**, laying a good foundation for building classification models. With lessons covering **logistic regression**, **classification metrics**, **model improvements**, **tree-based classification methods**, **support vector machines**, **k-nearest neighbours**, **naive Bayes classifiers**, **hyperparameter tuning**, and **neural network classifiers**, we've explored a wide array of classification algorithms. Using hands-on exercises and practical implementations, we've honed our skills in model construction, evaluation, and refinement.

By mastering these techniques, we're now equipped to tackle **real-world classification challenges** with **confidence** and **precision**.

What's next?

With the knowledge gained so far in **natural language processing (NLP)** and **classification**, we are well-equipped to **tackle a wide range of real-world data analysis tasks**. We can effectively preprocess text data, extract meaningful features, and build robust classification models.

Looking ahead, the next step to enrich our data analysis toolkit is to delve into **unsupervised learning techniques** which will provide powerful tools for uncovering hidden patterns and structures in **unstructured data**. We will continue to refine our analytical abilities and expand our toolkit continuing with our commitment to lifelong learning.

An illustration of two people, a man and a woman, celebrating the completion of a course. The man on the left is wearing an orange t-shirt and purple shorts, with his arms raised in a celebratory gesture. The woman on the right is wearing a blue long-sleeved shirt and purple pants, also with her arms raised. They are surrounded by a light blue circular area with scattered orange and blue confetti. The text "You've completed: NLP and classification" is centered in the middle of the illustration.

You've completed:
NLP and classification