**Week 4: (23/10/23)**

**Reading:**

* Machine learning with R: Chapter 14 (p494 -500 then 502 - 507)

Combining and managing the predictions of multiple models falls into winder set of meta-learning methods: techniques that involve learning how to learn. Ensembling – the task of modelling relationship between predictions of several models and desired outcome. Each classifier must be making an independent prediction, but also be doing more than guessing. Modes of increasing ensemble’s diversity: using assorted learning algorithms, manipulating training sample by taking different samples at random, manipulating single learning algorithm by using different hyperparameters, changing target feature representation, showing outcome as binary, categorical, numerical, partitioning training data into subgroups that represent different patterns to be learnt. Main strategies for composing final predictions: **Weighting methods** – gives score to each prediction on how heavily it will factor into final predictions, **Model stacking technique** – training secondary machine to make final prediction.

Benefits of ensembles:

* Use of independent ensembles allows work in parallel. - more rapid iterations and increase creativity.
* Improved performance on massive/ small datasets – fed subsets of features which is more computationally efficient.
* Ability to synthesise data from distinct domains – has ability to incorporate from multiple types of learners.
* More nuanced understanding of difficult learning tasks – split tasks into smaller portions so more likely to find patterns.

Bagging – generates new datasets on original training data, then used to generate models using single learning algorithms. Used with relatively unstable learners (models change when input data changed slightly) which is essential for ensuring diversity: Decision trees.

Random Forests – like bagging but adds diversity to decision tree by only allowing algorithm to choose from randomly selected subset of features each time it attempts to split.

Strengths: all-purpose model perform well on most problems, handles noisy or missing data and categorical or continuous features, selected only most important features, used on data with extremely large number of features.

Weakness: unlike DT, model is not easily interpretable, may struggle with categorical features with large numbers of levels, can’t be extensively tunes if greater performance desired.

* Machine learning with R: Chapter 6 (p175 – 179)

Regression involves specifying relationship between one numeric dependent variable (value being predicted) and one or more numeric independent variables (predictors). Regression methods used for statistical hypothesis testing, which determine if premise is likely to be true of false considering observed data.

Basic linear regression models (straight lines): simple linear regression has only 1 independent variable, multiple regression has 2 or more independent variables. Logistic regression used to model binary categorical outcome, Poisson regression models integer count data, multinomial Logistic regression models categorical outcome. These are all generalised linear models (GLM), these adapt straight lines to allow modelling of other forms of data.

Simple linear regression model defines relationship between dependent variable and single predictor variable using a line defines by y = a + Bx.

A(alpha): line crosses y axis

B(beta): change in y given increase of x.

**Regression**

2, 3, 12, 29, 42, 43, 53, 54, 76, 77, 78, 80, 83, 86, 88, 92, 94, 95, 98, 101, 102, 104, 105, 107, 109, 121, 124, 125, 129, 130, 132, 133, 138, 140, 141, 143, 144, 145

**Week 5: (30/10/23)**

**Reading:**

* Machine learning with R: Chapter 6 (p187 – 191)

GLM allows target to be non-normally distributed, non-continuous variable, allows variance of target variable to be related to its mean.

2 key components of any GLM:

1. Family refers to distribution of target features, must be chosen from members of exponential family of distribution. Chosen distribution may be discrete/continuous and may span different ranges of values.
2. Link function transforms relationship between predictors and target so it can be modelled using linear equation despite relationship being originally non-linear. Canonical link function, determined by chosen family and used by default or may choose different link to obtain better model fit.

GLM use maximum likelihood estimation (MLE), which finds parameter values for the specified distribution that are most likely to have generated the observed data.

* Machine learning with R: Chapter 7 (p229 – 240)

ANN uses a network of nodes to solve challenging learning problems.

Activation function is the mechanism which the artificial neuron processes incoming information and determines if it should pass signal to other neurons int the network. Threshold activation function – results in output signal only when specified input threshold has been attained. Unit step activation function – graph resembles a stair, outputs in binary so either 1 or 0.

Squashing function – squeezes input values into a smaller range of outputs, e.g., activation functions like sigmoid. Solution is to transform all inputs, so the feature values fall within a small range around zero.

Network topology is the number of layers, number of nodes and whether signals can travel backwards.

Input nodes receives unprocessed signals directly from input data. Each input node is responsible for processing single feature in dataset; then transformed by node’s activation set. Signal sent to output nodes which use activation function to generate final prediction. Input and output nodes are arranged in groups known as layers. Single layer networks are used for basic pattern classification.

Multilayer networks add one or more hidden layers that process signal from input nodes prior to output nodes. Hidden layers are why ANN are a black box model. A network with multiple hidden layers is a Deep Neural Network and training these networks is Deep Learning.

Simple multilayer networks are fully connected so every node in one layer is connected to every node in the next layer. Larger deep learning networks are partially connected. Removing connected limits amount of overfitting that occurs inside hidden layers. Networks where input signal is fed continuously in one direction from input layer to output layer are called feedforward networks. Recurrent neural network (RNN)/ feedback network allow signal to travel backwards using loops. Addition of short-term memory, or delay, increases power of recurrent networks immensely. Long short-term memory (LSTM) adapt model to have substantially longer recall.

Algorithms using strategy of back-propagating errors is known as backpropagation.

Strengths: can be adapted to classification/numeric prediction problems, makes fewer assumptions about data’s underlying relationships.

Weakness: computationally intensive and slow to train, prone to overfitting training data.

Backpropagation algorithm iterated through many cycles of two processes. Each cycle is known as epoch. Epoch includes:

* Forwards phase – input to output.
* Backward phase – networks output signal resulting from forward phase is compared to true target value in training data. Difference between output signal and true value results in an error is propagated backward in network to modify connection weights to reduce future errors. Gradient descent: backpropagation algorithm uses the derivative of each neuron's activation function to identify gradient in direction of each of incoming weights. Algorithm will change weights that result in greatest reduction in error by an amount known as learning rate. Greater learning rate, faster the algorithms will attempt to descend gradients.
* Machine learning with R: Chapter 7 (p251 – 256)

Support Vector Machine (SVM) can be imagined as a surface that creates a boundary between points of data plotted in multidimensional space representing examples and their feature values. Goal is to create flat boundary called hyperplane, which divides space to create homogeneous partition on either side. SVM learning combines aspects of instance based Nearest neighbour and linear regression modelling. This combination allows SVM to model highly complex relationships.

Linearly separable – separated perfected by a straight line or flat surface. Maximum margin hyperplane (MMH) creates the greatest separation between 2 classes. Maximum margin will improve chance that even if random noise is added, each class will remain on its own side of boundary.

Support vectors are the points from each class that are the closes to the MMH. MMH must have at least 1 support vector, which provide a compact way to store classification model.

Linearly separable data: MMH is far away as possible from outer boundaries of 2 groups of data. Outer boundaries are known as convex hull. MMH is perpendicular bisector of shortest line between the 2 convex hulls.

Nonlinearly separable data: Slack variable creates a soft margin that allows some points to fall on the incorrect side of the margin.

SVM has ability to maps problem into higher dimension space using process known as kernel trick, nonlinear relationship may appear linear. Involves process of constructing new feature that express mathematical relationships between measured characteristics.

Strengths: used for classification or numeric prediction problems, not overly influenced by noisy data and not prone to overfitting.

Weakness: slow to train, finding best model requires testing of various combinations of kernels and model parameter.

**Logistic Regression Classification**

6, 10, 11, 12, 14, 25, 26, 35, 36, 37, 39, 55, 74, 76, 83, 84, 85

**Artificial Neural Networks I**

4, 8, 14, 16, 17, 18, 19, 20, 21, 28-32, 33, 55-57, 59, 61-63

**Support Vector Machines**

3, 4, 5, 7, 9, 11, 13, 17, 26, 37, 39, 40, 46

**Week 6: (06/11/23)**

**Reading:**

* Machine learning with R: Chapter 15 (p526 – 540)

A deep nueral networks (DNN) is a neural network with more than one hidden layer. DNN’s use complex designed topologies to facilitate learning on big data, therefore capable of human-like performance on challenging learning tasks.

A convolutional neural network (CNN) is a deep feed-forward network used for visual tasks that independently learns the important distinguishing image features rather than requiring such feature engineering beforehand.

Topology of CNN:

* + - Convolutional layers are placed early in network and usually compromise the most computationally intensive step in network because they are the only layers to process the raw image data directly.
    - Pooling layers, also known as downsampling or subsampling layers. Gathers the output signals from a cluster of neurons in one layer and summarises them into a single neuron for the next layer, usually by taking maximum/average value among those being summarised.
    - Fully connecter layers are much like layers in traditional multilayer perceptron, used near the end of the CNN to build model that makes predictors.

The early stages of the networks perform feature engineering, while later stages use the constructed features to make predictions.

Transfer learning – sharing computational resources. (sharing pretrained neural networks).

Fine tuning – if knowledge doesn’t transfer directly to new task, it’s possible to hone a pre-trained neural network using additional training.

**Artificial Neural Networks II**

5, 10, 11, 15, 18, 19, 23 – 25, 27 – 30, 32, 35, 38, 42, 44, 45, 47

**Feature Extraction I**

2, 14, 15, 22, 42, 43, 44, 54, 61, 67, 99, 104, 121, 127, 128, 131, 145, 146, 151, 164, 173, 174, 175, 178, 188, 189, 190

**Week 7: (13/11/23)**

**Reading:**

* Computer Vision: Chapter 6 [(p370-372) (p375-386)]
* Machine learning with R: Chapter 9 (p242 – 307)

Clustering is an unsupervised machine learning task that automatically divides the data into clusters, or groups of similar items. Used for knowledge discovery rather than predictions because we do not tell the machine what we are looking for. It provides and insights into natural groupings found within data. Results in actionable data structures, which reduce complexity and provide insight into patterns of relationships.

Clustering is to create new data. Unlabelled examples are given a new cluster label, which is inferred from relationships within the data. Clustering tasks are referred to as unsupervised classification as it classifies unlabelled examples.

Various algorithms are distinguished by 2 main characteristics:

* Similarity metric, which provides quantitative measure of how closely two examples are related.
* Agglomeration function, which governs the process of assigning examples to clusters based on similarity.

Three main clusters of clustering algorithms (simplest -> sophisticated)

1. Hierarchical methods – which creates a family tree-style hierarchy that positions the most similar examples more closely in graph structure (dendrogram). Doesn’t have a predefined expectation of clusters, can have one giant cluster, clusters of just one and varying cluster sizes. As the number of examples grows, the number of calculations to compute the similarity grows (N\*N = N^2) and increase in memory to store result in similarity matrix, therefore only used for small datasets.
2. Partition-based methods – which treats the examples as points in multidimensional space and attempt to find boundaries in this space that led to relatively homogenous groups, requires fewer comparisons between examples so quite computationally efficient. It is somewhat rigid/ random when it comes to group assignments. If an example falls on a boundary it will be placed randomly into one cluster or the other.
3. Model or density-based methods – which rely on statistical principles/ density of points to discover fuzzy boundaries between clusters. Estimates probability that an example belongs to each cluster.

Mixture modelling attempts to disentagle datasets composed of examples pulled from a mixture of statistical distributions. It requires knowledge of how many distributions are involved but also an assumption of the types of distributions. This may be too rigid for real-world clustering tasks.

DBSCAN “density-based spatial clustering of applications with noise” approach used to identify natural clusters in data. A process of creating neighbourhoods of examples that are all within a specified radius of other examples in the cluster. Predefined number of core points within radius forms initial cluster nucleus, points within radius added to cluster and compromise outmost boundary. Any points not close to core points treated as noise.

Agglomerative clustering begins with each example in its own cluster, then connects most similar examples first until all examples are connected in a single cluster.

Divisive clustering begins with a single large cluster and ends with all examples in their own individual clusters.

K-means algorithm is an example of partition-based clustering. Assigns one of the n examples to one of the k clusters, where k is a number that has been determined ahead of time. The goal is to minimise the differences in feature values of examples within each cluster and maximise the difference between each cluster. Locally optimal – starts with initial guess for cluster assignments, then modifies assignment slightly to see for changes that improve homogeneity within clusters. Stops when there is no change in reassignment of examples. Elbow methos attempts to measure how the homogeneity or heterogeneity within clusters changes for various values of k.

Strengths:

* Uses simple principes that can be explained in non-statistical terms.
* Highly flexible and can be adapted with simple adjustments to address many of its shortcomings.
* Performs well enough under many real-world use cases.

Weaknesses:

* Not as sophisticated as more modern clustering algorithms.
* Because it uses an element of random chance, it’s not guaranteed to find optimal set of clusters.
* Requires reasonable guess as to how many clusters naturally exist in the data.
* Not ideal for non-spherical clusters or clusters of widely varying density.

**Feature Extraction II**

25, 27, 30, 86, 87, 91, 93 - 102

**Unsupervised Learning, I**

21, 36, 56, 57, 58, 64, 66, 82, 85, 86, 87

**Week 8: (20/11/23)**

**Reading:**

* Machine learning with R: Chapter 13 (p446)

A possible way to reduce the dimensionality of a highly dimensional dataset is to synthesis a smaller number of composite predictors. This is the goal of feature extraction, a dimensionality reduction technique that creates new features rather than selecting a subset of existing features. Extracted features are constructed so they reduce amount of redundant information and keep useful information.

**Week 9: (27/1/23)**

**Reading:**

[16. Natural Language Processing with RNNs and Attention | Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition (oreilly.com)](https://learning.oreilly.com/library/view/hands-on-machine-learning/9781098125967/ch16.html#idm45720177136448)

* Chapter 16
* Chapter 15