Machine Learning

P. S. Langeslag

Rule-Based Algorithm

A cluster of hand-written `if ... then` rules.

The default method for a range of NLP tasks, such as

- Lemmatization
- Named-entity recognition (NER)

Supervised Machine Learning

Training an algorithm on hand-labelled data until it correctly handles new data.

Typical applications include classification (“select all squares with traffic lights”).

Unsupervised Machine Learning

Training a classifier on unlabelled data.

Common NLP applications:

- topic modelling
- word vectorizing

Symbolic vs Subsymbolic AI

Symbolic AI

Learning algorithms relying on explicit, human-readable logic.

Subsymbolic AI

Primarily neural networks, whose learning process relies on calculations, thresholds, and implicit weight adjustments on lower-level features that cannot be so straightforwardly represented as human-readable logic.

Naive Bayes Classifier

Probabilistic model assigning scores for individual features without considering probabilities following from their correlation.

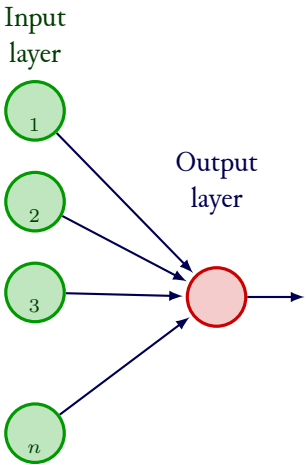
Naive Bayes Classifier

Probabilistic model assigning scores for individual features without considering probabilities following from their correlation.

The model works acceptably well, and it allows for the training of a classifier on limited training data.

Naive Bayes is a **batch learning** model, i.e. it does not adjust weights between learning cycles (there are no cycles); it finds patterns in the full set of data once, then uses this database of patterns to classify any new data it is fed.

Perceptron



Algorithm consisting of a **single node** processing input into output and autonomously adjusting the weighting of each input after each training cycle until the actual output *just* matches the expected output.

- Supervised learning
- Binary classification (outputs `0` or `1`)
- Linearly separable data

(See Lane et al. ch. 5.)

Naive Bayes vs Perceptron Classifiers

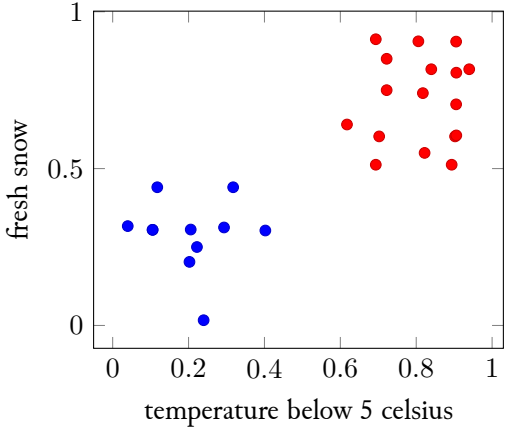
Naive Bayes

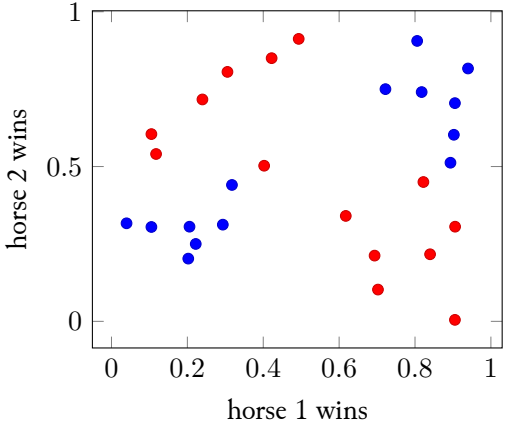
- Batch learning: observe all data, then set parameter values.
- Generative model: $P(X, Y)$ (the probability of X and Y both being true).
- Think coin flips: we calculate the probability of a "series" of inputs.

Perceptron

- Online learning: parameter values are updated after every cycle.
- Discriminative model (conditional probability): $P(Y|X)$ (the probability of Y given X).
- Alters the weights of its inputs for each cycle until its function just accounts for the labels on your data.

Linearly vs Nonlinearly Separable Data: The XOR Problem





(See Lane et al. pp. 164–165; TensorFlow Playground.)

Backpropagation

Backpropagation of the error calculates how much a particular weight contributed to the overall error in a training cycle (**epoch**) and adjusts the weights accordingly before running the next cycle.

In a multilayer neural network, this requires nonlinear maths, because we can't see the outputs of any but the final layer of neurons.

(See Lane et al. ch. 5.)

Artificial Neural Network (ANN)

Algorithm relying on **multiple nodes** processing input and autonomously adjusting the weighting of each input after each training cycle until the actual output *just* matches the expected output.

The distinction with a perceptron is that backpropagation requires a nonlinear activation function and therefore multiple nodes (to model nonlinear relationships between input and output).

Artificial Neural Network (ANN)

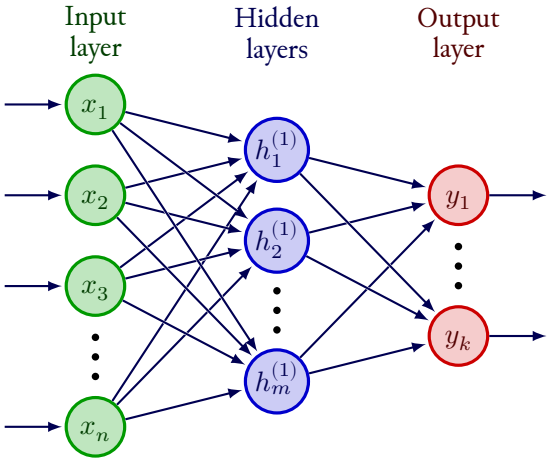
Algorithm relying on **multiple nodes** processing input and autonomously adjusting the weighting of each input after each training cycle until the actual output *just* matches the expected output.

The distinction with a perceptron is that backpropagation requires a nonlinear activation function and therefore multiple nodes (to model nonlinear relationships between input and output).

- Supervised and unsupervised learning
- Linearly and nonlinearly separable data
- Classification, regression analysis, clustering, filtering, etc.

(See Lane et al. ch. 5.)

Deep Learning



A neural network relying on multiple (layers of) neurons, allowing nonlinear classification.

Overfitting

Training a supervised neural network so precisely on its training data that its ability to predict new data is adversely affected.

Data Separation

- fitting

- selecting optimal hyperparameters

- demonstrating accuracy with new data

Shuffle your data!

Bibliography

Bird, Steven, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly, 2009. https://www.nltk.org/book/.

Eisenstein, Jacob. *Introduction to Natural Language Processing*. Cambridge, MA: MIT Press, 2019.

Jurafsky, Dan, and James H. Martin. *Speech and Language Processing*. 3rd ed. draft., 2021. http://web.stanford.edu/~jurafsky/slp3/.

Lane, Hobson, Cole Howard, and Hannes Hapke. *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text with Python*. Shelter Island, NY: Manning, 2019.

Matthes, Eric. *Python Crash Course*. 2nd ed. San Francisco, CA: No Starch, 2019.

Python Software Foundation. "Python," October 4, 2020. https://www.python.org/.

Vasiliev, Yuli. *Natural Language Processing Using Python and spaCy: A Practical Introduction*. San Francisco: No Starch Press, 2020.