# Introduction for Machine Learning

## Important Point: In machine learning and pattern recognition, a feature is an individual measurable property or characteristic of a phenomenon being observed. Choosing informative, discriminating and independent features is a crucial step for effective algorithms in pattern recognition, classification and regression.

# NumPy:

Numpy is one of the fundamental packages for sceintific computing in Python. It contains functionality for multidimensional arrays, high level mathmetical functions such as linear algebra operations and fourier transform(The Fourier transform decomposes a function into its constituent frequencies), and pseudorandom number generation.

* Scikit-learn takes in data in the form of NumPy arrays.

# SciPy:

SciPy is a collection of functions for sceintific computing in Python. It provides advanced linear algebra.

## scikit-learn draws from SciPy’s collection of functions for implementin its algorithms.

## The most important part of SciPy for us is scipy.sparse: this provides spars matrics, which are another representation that is used for data in scikit-learn

## Imoportant point: Saprse matrics are used whenever we want to store a 2D array that contains mostly zeros.

# Sparse martrics:

**Sparse matrices** are memory efficient data structures that enable us store large **matrices** with very few non-zero elements **aka sparse matrices**. In addition to efficient storage, **sparse matrix** data structure also allows us to perform complex **matrix** computations.

# CSR or The compressed sparse row

The compressed sparse row (**CSR**) or compressed row storage (CRS) or Yale **format** represents a matrix M by three (one-dimensional) arrays, that respectively contain nonzero values, the extents of rows, and column indices.

thinly dispersed or scattered.

## Sparse means: scattered.

# COO or The coordinate list:

Coordinate (of the same rank; equal) list (**COO**) **COO** stores a list of (row, column, value) tuples. Ideally, the entries are sorted first by row index and then by column index, to improve random access times. This is another **format** that is good for incremental matrix construction.

# Matplotlib

matplot lib is the primary scientific library in Python. it provides functions for making publication-quality visualizations such as line charts, histograms,scatter polts.

## Important note: Bunch object is very similar to dictionary, it contains keys and values.

## Important note: The individual items are called *samples* in machine learning, and their properties are called *features*.

## Important note: The *shape* of the data array is the number of *samples* multiplied by the number of *features*. This is a convention in scikit-learn, and your data will always be assumed to be in this shape.

## Important note: We cannot use the date we used to build the model to evaluate it. Why? Because our model can always simple remember the whole training set, this remembering does not indicate to us whether it will also perform well on new data.

## Important note: Using a test set containing 25% of the data is good rule of thumb.

## Important note: In scikit-learn, data is usually denoted with a captial X, while labels are denoted by a lower case y. Why? We use X because the data is a two-dimensional array (a matrix) and a lowercase y because the target is a one-dimensional array(a vector).

# What is random\_state parameter?

To make sure that we will get the same output if we run the same function several times, we provide the pseudorandom number generator with a fixed seed using the random\_state parameter. This will make the outcome deterministic, so this line will alwayse have the same outcome.

## Pseudo Random Number Generator(PRNG) refers to an algorithm that uses mathematical formulas to produce sequences of random numbers. PRNGs generate a sequence of numbers approximating the properties of random numbers. A PRNG starts from an arbitrary starting state using a seed state.

## Important Point: A random seed (or seed state, or just seed) is a number (or vector) used to initialize a pseudorandom number generator. For a seed to be used in a pseudorandom number generator, it does not need to be random

# What is the first step before buiding a machine learning model?

Inspect the data with scatter plot.

## Create a dataframe from data in X\_train.

## label the column using the strings in the original data set.

## Create a scatter matrix from the dataframe, color by y\_train

# What is the Estimator classes?

All machine learning models in scikit-learn are implemented in their own classes, which are called **Estimator classes**.

## Important Point: The K-nearest neighbors classification algorithm is implemented in the KneighborsClassifier class in the neighbors module.

## Important Point: Before we can use the model, we need to instantiate the class into an object. This is when we will set any parameters of the model.

## Important Point: Scikit-learn always expects two-dimensional arrays for the data.

# Name a situation that the test set will comes in, why?

In evaluation the method, because we know the right items are in the test set. So, we can make a prediction for each iris in the test data and compare it against its label.

## Important Point. what lable means in the current example? THE KNOWN SPECIES.

# How can we measure how well the model works?

By computing the accuracy, which is the fraction of folwers for which the right species was predicted.

## Important Point. So!! Calculating the accuracy means: the fraction from the right labels(here the right species).

# When we use supervised learning?

Whenever we want to predict a certain outcome from a given input, and we have examples of input/output pairs.

## the goal is making predictions for new, never-before-seen data.

## Important Point: supervised learning often requires human effort to build the training set, but afterward automates and often speeds up an otherwise laborious task.

# What is Unsupervised algorithms?

In unsupervised learning, only the input data is known, and no known output data is given to the algorithm.

## Important Point: In binary classification ‘positive’ is the object of the study.

## Important Point: The goal of regression in tasks is to predict a continuous number, or a floating-point number in programming terms (or real number in mathematical terms).

## Important Point: If a model is able to make accurate predictions on unseen data, we say it is able to generalized from the training set to test set.

## The only measure of whether an algorithm will perform well on the new data is the evaluation on the test set.

# What overfitting means?

Overfitting occures when you fit the model too closely to the particularities of the training set and obtain a model that works well on the training set but is not able to generalize to new data.

## Choosing too simple a model is called underfitting.

# When we can import and instantiate the class?

When we can set parameters, like the number of neighbors to use.

# K-Nearest Neighbors:

## Arguably the simplest learning algorithm.

## building the model consists only of the storing the training dataset.

## To make a predicition for a new data point, the algorithm finds the closest data point in the training dataset-its “nearest neighbors”.

## In the simplest version, the K-NN algorithm only considers exactly one nearest neighbor, which is the closest training data point to the point we want to make a prediction for.

## the prediction is then simply *the known output* for this training point

## Important Point: The prediction of the one-nearest-neighbor algorithm is the label of that point.

## Important Point: Where the K-nearest neighbors algorithm came from? Instead of conidering only the closest neighbor, we can consider an arbitrary number, k, of neighbors.

## Important Point:When we considering more than one neighbor, we use voting to assign a label. This means for each test points, we count how many neighbors belong to class 0 and how many neighbors belong to class 1. We then assign the class that is more frequent: in other words, the majority class among the k-nearest neighbors.

## Important Point: Forge is an example of a synthetic two-class classification dataset, which has two features.

# What is mglearn?

Mglearn is a library of utility functions we wrote for the book of Introduction to ML.

# How we can apply K-nearest-neighbors algorithm using scikit-leanrn?

1. Split our data into train and test set so we can evaluate generalization performance.
2. We import and instantiate the class. This is when we can set parameters, like the number of neighbors to use.
3. We fit the classifier using the training set. (For KneighborsClassifier this means storing the datasets, so we can compute neighbors during prediction.
4. We call the **predict method**, which for every data point in the data set, this will compute its nearest neighbors in the training set and finds the most common class among them.
5. To evaluate how well our model generalizes, we can call the **score method**

with the test together with the test data together.

## Important Point: What we can figure out from the results? We can see the model is x% accurate, meaning the model predicted the class correctly for x% of the samples in the test dataset.

# What is decision boundry?

When we color the plane according to the class that would be assigned to a point in that region, this lets us veiw the decision boundry, which is the divide between where the algorithm assigns class 0 versus where it assigns class 1.

## Important Point: Using few neighbors corresponds to high model complexity.

## Importan Point: The prediction using a single neighbor is just the target value of the nearest neighbor.

## Important Point: When using multiple nearest neighbors, the prediction is the average, or mean, of the relevant neighbore.

## The k-nearest neighbors algorithm for regression is implemented in the Kneighbors Regressor class in scikit-learn, and it’s used similarly to KneighbborsClassifier.

## Important Point: For evaluate a regresson model we can use the R2 score as the coefficient of determination, which is a measure of goodness of a prediction for regression model, and yields a score between 0 (Constant model that predicts the mean of the training set responses) and 1(perfect prediction).

## Important Point: In principle, There are two important parametes to the Kneighbors classifier: the number of neighbors and how you measure distance between data points. By default, Euclidean distance is used, which works well in many settings.

## Kneighbors does not perform well on datasets with many features(hundreds or more), and it does particularly badly with datasets where most features are 0 most of the time (so-called sparse datasets).

## Important Point: Linear Regression is a machine learning algorithm based on supervised learning. ... Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).

# What are Linear Models?

1. Linear models make a prediction using a linear functions of the input features.
2. For regressaion, the general prediction formula for a linear model looks as: ŷ = w[0] \* x[0] + w[1] \* x[1] + ... + w[p] \* x[p] + b
3. w an b are parameters of the model that are learned, and ŷ is the prediction the model makes.
4. Alternatively, you can think of the predicted response as being **a weighted sum of the input features**, with weights given by the entries of w.

## 

## Important point: In Linear Regression the intercept is where the prediction line (y) should cross the y-axis.

## REGRESSION: In statical modelling, regression analysis is the process of estimating the relationship between a dependent variable (outcome variabel) and some independent variables(predictors, covariates, features).

## LINEAR REGRESSION OLS(Ordinary least sqaures): The researcher finds a line that fits the data according to a specific mathematical criterion.

## Important point: Criteria is the plural form of criterion. It is used when referring to more than one criterion. Criterion is singular and is used to refer to a single thing.

## In geometry, a hyperplane is a subspace whose dimension is one less than that of its ambient space. If a space is 3-dimensional then its hyperplanes are the 2-dimensional planes, while if the space is 2-dimensional, its hyperplanes are the 1-dimensional lines.

# Linear Regression, OLS(aka Ordinery least squares):

1. Linear regression finds the parameters w and b that minimize the mean squared error between predictions and the true regression targets, y, on the training set.
2. The mean square is the sum of the squared differences between the predictions and the true values.
3. the slope parameters(w), also called weights or coefficients, are stored in the coef\_attribute.
4. the offset or intercept(b) is stored in the intercept\_attribute.

## Important Point: The discrepancy between performance on the training set and the test set is a clear sign of overfitting, and therefore we should try to find a model that allows us to control complexity.

## Reminder!!! OLS or Ordinary Least Square doesn’t have any parameters, which is a benefit, but it also has no way to control the complexity.

# The Ridge Regression:

1. The Ridge regression is also a linear model for regression, so the formula it uses to make predictions is the same one used for ordinary least squares.
2. In Ridge regression, the coefficients(w) are chosen to:
   * 1. Predict well on the training data.
     2. Fit an additional constraint.
3. BUT!!!!! In Ridge regressions the magnitude of coefficients to be as small as possible; MEANING!!!! all entries of w should be close to zero, MEANING!!!! Each feature should have as little effect on the outcomes as possible = Having small slope.(This constraint is an example of what is called regularization.)
4. alpha parameter: since the ridge model makes a trade-off between the simplicity and its performance on the training set, the user can specify the amount of importance between these two using the alpha parameter.
   * 1. THE OPTIMUM setting of alpha depends on the particular dataset we are using.
     2. Increasing alpha, forces coefficients to move more toward zero, which decreases training set performance but might help generalization.
5. Mathematically, Ridge penalizes the L2 norm of the coefficients, or the Euclidean length of w.
   * 1. In mathematics, the Euclidean distance or Euclidean metric is the "ordinary" straight-line distance between two points in Euclidean space.

## Important Point: In mathematics, statistics, and computer science, particularly in machine learning and inverse problems, regularization is the process of adding information in order to solve an ill-posed problem or to prevent overfitting. Regularization applies to objective functions in ill-posed optimization problems.

## Importan Ponit: With Linear Regression we overfitting. Ridge is a more restricted model, so we are less likely to overfit.

## Reminder: A less complex model means worse performance on the training set, but better generaliztion. SO!!!! as we are only interested in generaliztion performance, we should choose the Ridge model over the linear Regression.

## Plots that show model performance as a function of dataset size are called learning curve.

# Lasso

1. Lasso is an alternative to Ridge for regularizing linear regression.
2. Lasso also restricts coefficients to be close to zero, but by L1 regularization: The result of using L1 is: some coefficients are exactly zero.
   * 1. MEANING!! Some features are entirely ignored by the model, MEANING!! This can be seen as a form of automatic feature selection.
     2. Having some coefficients exactly zero often makes a model easier to interpret, and can reveal the most important features of your model.
3. The lasso penalizes the L1 norm of the coefficient vector, or in other words, the sum of the absolute values of the coefficient.

## Important Point:  In machine learning “penalization” is used to avoid/ lessen the overfitting of the model.

1. If you have a large amount of features and expect only few or them to be important Lesso might be a betther choice.
   1. and also if you need a model that is easier to understand, as it will select only a subset of the features.

# Linear models for classifications:

1. Prediction formula:

ŷ = w[0] \* x[0] + w[1] \* x[1] + ... + w[p] \* x[p] + b > 0

* + 1. This formula is very similar to linear regression, but instead of just returning the weighted sum of the features, we threshold the predicted value at zero.

## Reminder!! In CLASSIFICATION the goal is to predict a class label, which is a choice from a predefined list of possibilities.

1. If the function is smaller than zero, we predict the class is -1; if it is larger than zero, we predict the class is +1.

## Imoprtant Ponit: This prediction is very common in all linear models for classification.

1. For linear models for regression, the output, ŷ, is a linear function of the features: a line, a plane, a hyperplane. So!! for linear models for classification, the decision boundary is a linear function of the input. MEANING!! a (binary) linear classifier is a classifier that seprates two classes using a line, a plane, or a hyperplane.

## Reminder: The intercept is where the prediction line should cross the y-axis.

1. There are many algorithms for learning linear models, these algorithms all differ in the following two ways:
   * 1. How well they measure a particular combination of coefficients and intercept fits the training data.
     2. If and what kind of regularization they use.

## Important Point: Machines learn by means of a loss function. It’s a method of evaluating how well specific algorithm models the given data. If predictions deviates too much from actual results, loss function would cough up a very large number. Gradually, with the help of some optimization function, loss function learns to reduce the error in prediction.

1. The two common linear classification algorithms are:
   * 1. Logistic Regression, implemented in linear\_model.LogisticRegression.
     2. Linear support vectore machines(linear SVMs), implemented in svm.LinearSVC (SVC stands for support vector classifier).
     3. for both the trade-off parameter that determines the strength of the regularization is called C, and higher values of C correspond to less regularization.
     4. When you use a high value of the parameter C, logicticRegression and LinearSVC try to fit the training set as best as possible, while with values of C, the model put more emphasis on finding a coefficient vector that is close to zero.
     5. Using low values of C will cause the algorithms to try to adjust the “majority” of data points, while using the high value of C stresses the importance that each individual data point be classified correctly.

## Important Point: Similarly to the case of regression, linear models for classification might seem very restrictive in low-dimensional spaces, only allowing for decision boundaries that are straight lines or planes.

## In High dimensions, linear models for classification become very poweful, and guarding against overfitting becomes increasingly important when considering more features.

## Reminder: Choosing very simple a model is called underfit: When training and set performance are very close, it is likely that we are underfitting.

## REMINDER: What a simple model means? When the coefficients of the model is near to zero.SO!! When we increased alpha, this will force the coefficients to move more toward zero, which increases training set performance BUT!! might help generalization.

# What are the main difference between the models in Regression?

The main difference is the penalty parameter.(Penalty Parameter influences on the regularization and wheter the model will use all available features or select only a subset.)

## Important point: penalty parameter influence the regularization and whether the model will use all available features or select only a subset.

# What is *one-vs-rest* approach?

1. A common technique to extend a binary classification algorithm to multiclass classification algorithm.
2. In the one-vs.-rest approach, a binary model is learned for each class that tries to seprate that class from all of the other classes, resulting in as many binary models as there are classes.
3. To make a prediction, all binary classifiers are run on a test point. The classifier that has the highest score on its sinlge class “wins” and this class label is returend as the prediction.
4. Having one binary classifier per class results in having one vector of coefficients(w) and one intercept(b) for each class.

## Reminder: The intercept is where the prediction line should cross the y-axis, or b!

## Important Point: When graphed, a linear equation will have a straight line that has a constant slope. ... A linear equation will have constantly increasing y values and a straight line, while a nonlinear equation will have outputs increasing at a non-constant rate and a curved graph.

# What is Logarithmic Scale:

A logarithmic scale (or log scale) is a way of displaying numerical data over a very wide range of values in a compact way—typically the largest numbers in the data are hundreds or even thousands of times larger than the smallest numbers. Such a [scale](https://en.wikipedia.org/wiki/Scale_(measurement)) is [nonlinear](https://en.wikipedia.org/wiki/Nonlinear): the numbers 10 and 20, and 90 and 100, are not the same distance apart on a log scale. Rather, the numbers 10 and 100, and 100 and 1000 are equally spaced. Thus moving a set distance along the scale means the number has been multiplied by 10 (or some other fixed factor). Often exponential growth curves are displayed on a log scale, otherwise they would increase too quickly to fit within a small [graph](https://en.wikipedia.org/wiki/Plot_(graphics)). Another way to think about it is that the number of [digits](https://en.wikipedia.org/wiki/Numerical_digit) of the data grows at a constant rate. The numbers 10, 100, 1000, and 10000 are equally spaced on a log scale, but the number of digits is going up by 1 each time: 2, 3, 4, and 5 digits. In this way, adding two digits multiplies the quantity measured on the log scale by a factor of 100.

# When we should use linear models?

1. When we want to get fast results for training and predicting on large datasets.
2. When we have sparse data.
3. When the number of features is large compared to the number of samples.

# What is Naive Bayes Classifiers?

1. Naive Bayes Classifier are a family of classifiers that are quit similar to the linear models.
2. They tend to be faster in trainig.
3. Importan point about them: The price for this efficiency is that naive Byes Models often provide, slightly worse generalization performance.
4. The reason that naive Bayes models are so efficient is that they learn parameters by looking at each feature individualy and collect simple pre-class statistics from each feature.

**نکته قابل توجه:** معادله برنولی یا اصل برنولی در مکانیک سیالات رفتار [شاره](https://fa.wikipedia.org/wiki/%D8%B4%D8%A7%D8%B1%D9%87) را در [جریان یکنواخت](https://fa.wikipedia.org/wiki/%D8%AC%D8%B1%DB%8C%D8%A7%D9%86_%DB%8C%DA%A9%D9%86%D9%88%D8%A7%D8%AE%D8%AA) توضیح می‌دهد و فرم ریاضی [قانون بقای انرژی](https://fa.wikipedia.org/wiki/%D9%BE%D8%A7%DB%8C%D8%B3%D8%AA%DA%AF%DB%8C_%D8%A7%D9%86%D8%B1%DA%98%DB%8C) در [سیالات](https://fa.wikipedia.org/wiki/%D8%B3%DB%8C%D8%A7%D9%84%D8%A7%D8%AA) است. به زبان ساده چنین است: در [شاره‌ای](https://fa.wikipedia.org/wiki/%D8%B4%D8%A7%D8%B1%D9%87) که جریان دارد، افزایش سرعت جریان با کاهش فشار هم‌زمان است، به شرطی که ارتفاع سیال ثابت بماند. معادله برنولی بیان دقیق‌تر این اصل است، به عبارت دیگر اگر [سرعت](https://fa.wikipedia.org/wiki/%D8%B3%D8%B1%D8%B9%D8%AA" \o "سرعت) یک [سیال](https://fa.wikipedia.org/wiki/%D8%B3%DB%8C%D8%A7%D9%84) افزایش پیدا کند، [فشار](https://fa.wikipedia.org/wiki/%D9%81%D8%B4%D8%A7%D8%B1)ی که بر یک سطح وارد می‌کند کاهش می‌یابد و بالعکس.

1. There are three kinds of naive Bayes classifiers implemented in scikit-learn: GaussianNB, BernoulliNB, MultinominalNB
   * 1. Guassian can be applied to any continuous data.
        1. Stores the average value as well as the standard deviation of each feature for each class)
        2. Mostly used on very high dimensional data.
     2. Bernoulli assumes binary data (Counts how often every feature of each class is not Zero)
        1. Used in spare data such as text.
     3. MultinominalNB (Takes into account average value of each feature for each class.)
        1. Used in spare data such as text.
        2. MultinominalNB usually performs better than BinaryNB, particularly on datasets with a relatively large number of nonzero features(i.e., large documents).
2. Importan point coef\_ for the naive Bayes models has a somewhat different meaning than in the linear models, so the coed\_ is not the same as *w*.