Introduction to Data Science: SWE 227

Topics to be covered

- Classifiers: (Bayesian, Maximum a posteriori, parameter estimation, decision tree, SVM, bag of words, N-gram models, association rules, nearest neighbor, locally weighted regression)
- Clustering (mixture models, k-means clustering, hierarchical clustering, distributional clustering)
- Ensemble techniques.

1. Classifiers

Intro

The goal is to attempt to classify each observation into a category (aka, class or cluster) defined by Y , based on a set of predictor variables (aka, features), X.

- Bayesian
- Maximum a posteriori
- parameter estimation
- decision tree
- SVM
- bag of words
- N-gram models
- association rules
- nearest neighbor
- locally weighted regression

1. Classifiers

Bayesian

We defined conditional probability as:

- $P(B|A) = P(B \cap A) / P(A)$

And using the fact that $P(B \cap A) = P(A|B)P(B)$ we get
Bayes' Theorem:

- $P(B|A) = P(A|B)P(B) / P(A)$

Another version of Bayes' Theorem is found by substituting in the Law of Total Probability (LOTP) into the denominator:

- $P(B|A) = P(A|B)P(B) / [P(A|B)P(B) + P(A|B^C)P(B^C)]$

1. Classifiers

Bayesian (Lets escalate things)

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

$$X = (x_1, x_2, x_3, ..., x_n)$$

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

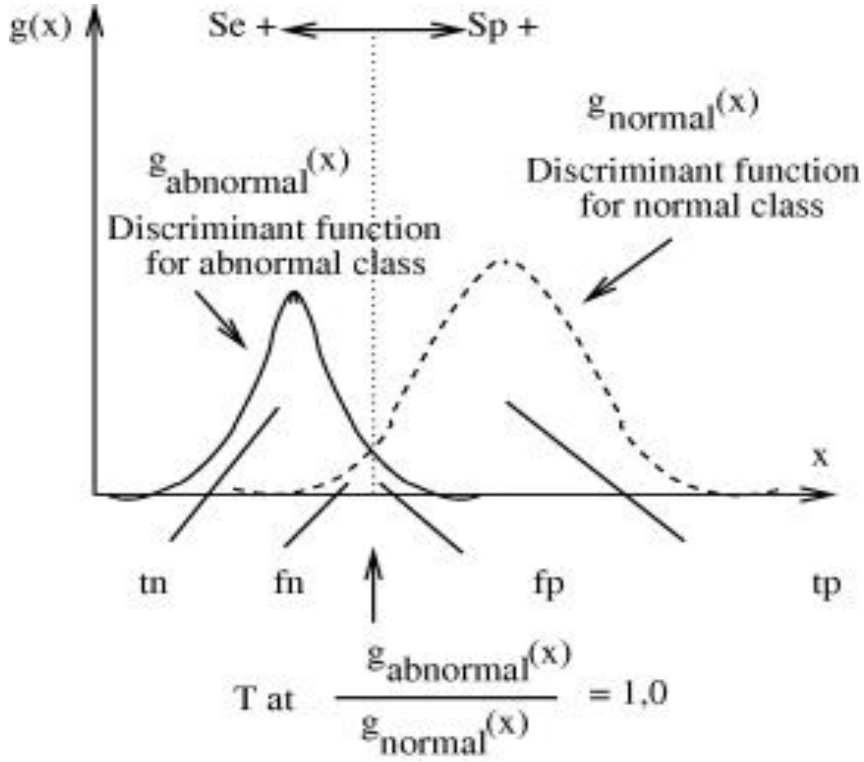
$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

[Dealing with Multivariates]

1. Classifiers

Bayesian (Categories and Decision Boundaries)

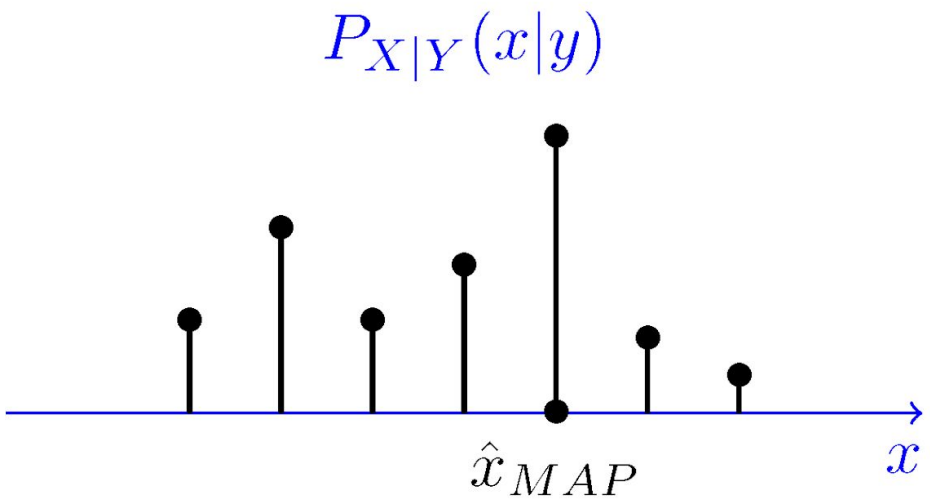
Multinomial Naive Bayes
Bernoulli Naive Bayes [bool features; x_i either 0 or 1]
Gaussian Naive Bayes [Continuous features; say 0 < x_i <1]

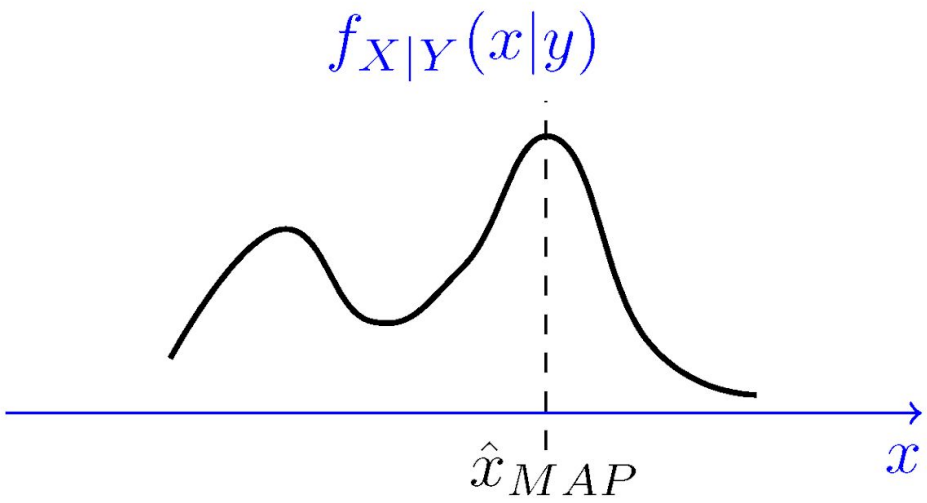


1. Classifiers

Maximum a Posteriori

In Bayesian statistics, a maximum a posteriori probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution.





1. Classifiers

Maximum a Posteriori

In Bayesian statistics, a maximum a posteriori probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution.

The MAP estimate of the random variable X, given that we have observed Y=y, is given by the value of x that maximizes

$f_{X|Y}(x|y)$ if X is a continuous random variable

$P_{X|Y}(x|y)$ if X is a discrete random variable.

The MAP estimate is shown by $\hat{x}_{MAP}$.

1. Classifiers

Maximum a Posteriori

Objective: Find Value of x that maximize $f_{X|Y}(x|y) = f_{Y|X}(y|x)f_X(x) / f_Y(y)$. [Ofcourse in a given limit.] Only maximizing the numerator would suffice though.

Let X be a continuous random variable with the following PDF:

$$f_X(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1 \\ \\ 0 & \text{otherwise} \end{cases}$$

Also, suppose that

$$Y \mid X = x \quad \sim \quad Geometric(x).$$

Find the MAP estimate of X given $Y = 3$.

1. Classifiers

We know that $Y \mid X = x \quad \sim \quad Geometric(x)$, so

$$P_{Y|X}(y|x) = x(1-x)^{y-1}, \quad \text{for } y = 1, 2, \cdots.$$

Therefore,

$$P_{Y|X}(3|x) = x(1-x)^2.$$

We need to find the value of $x \in [0, 1]$ that maximizes

$$\begin{aligned} P_{Y|X}(y|x) f_X(x) &= x(1-x)^2 \cdot 2x \\ &= 2x^2(1-x)^2. \end{aligned}$$

We can find the maximizing value by differentiation. We obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[x^2(1-x)^2\right] = 2x(1-x)^2 - 2(1-x)x^2 = 0.$$

Solving for x (and checking for maximization criteria), we obtain the MAP estimate as

$$\hat{x}_{MAP} = \frac{1}{2}.$$

1. Classifiers

Maximum a Posteriori (Home Task)

Let X be a continuous random variable with the following PDF
$f_X(x) = 4x^2$ when $0 <= x <= 4$ and 2 otherwise.

Suppose that $P_{Y|X}(y|x) = xe^y$ for $y < 0$

FInd the MAP estimate of X given Y = 2

1. Classifiers

Parameter Estimation

Parameter estimation is defined as the experimental determination of values of parameters that govern the system behavior, assuming that the structure of the process is known.

Parameter Estimation is a branch of statistics that involves using sample data to estimate the parameters of a distribution.

Some fantastic things that people have done using this:
https://www.sciencedirect.com/topics/earth-and-planetary-sciences/parameter-estimation

For more reading
http://www.fao.org/3/x8498e/x8498e0e.html

1. Classifiers

Parameter Estimation(Some Vocab)

If the value an estimator estimates for the parameter, θ', always converges to the actual parameter value θ as the quantity of data used for parameter estimation increases, we say an estimator is **consistent**.

The **bias** of an estimator is the deviation of the expectation from the actual true value. If, for a given estimator, the bias is zero, we say that that estimator is unbiased.

If an estimator has lower variance than another we say it is more **efficient**, and we can calculate the efficiency of estimator p relative to estimator q as (Var(θ'p))/Var(θ'q).

1. Classifiers

Decision Tree

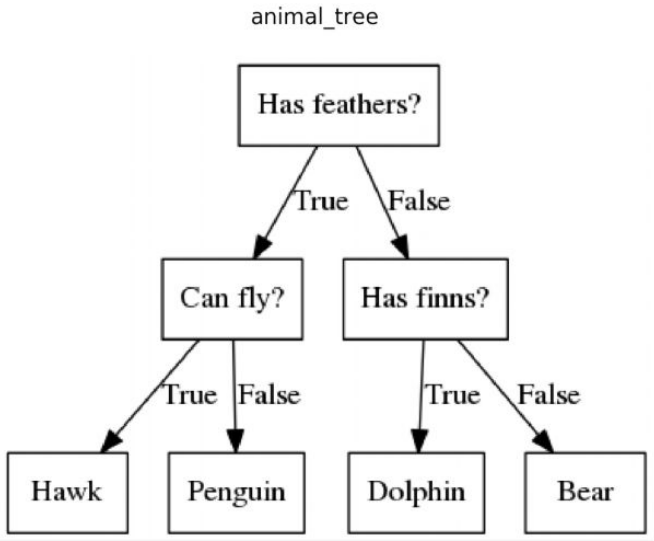
Motivation: Interpretability
Approach: Methodical and explainable

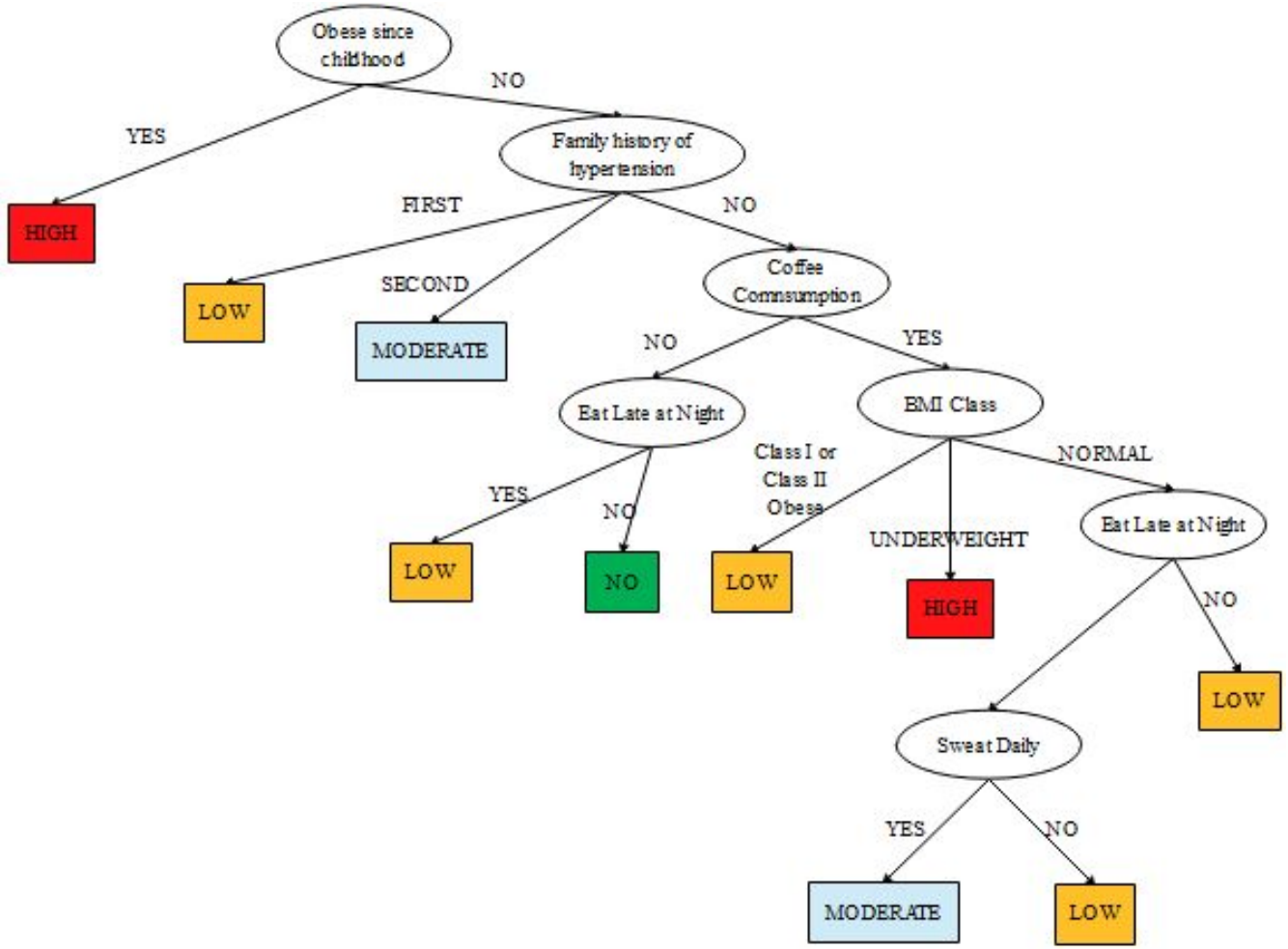
Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

Recursive partitioning: A tree can be “learned” by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions.

1. Classifiers

Decision Tree





1. Classifiers

Decision Tree (Strength)

Strength

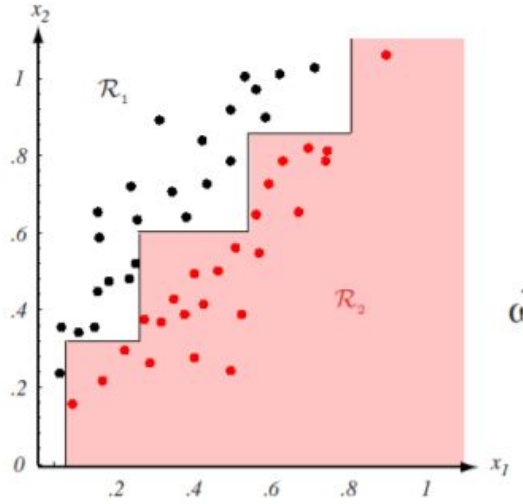
- ❏ Decision trees are able to generate understandable rules.
- ❏ Decision trees perform classification without requiring much computation.
- ❏ Decision trees are able to handle both continuous and categorical variables.
- ❏ Decision trees provide a clear indication of which fields are most important for prediction or classification.

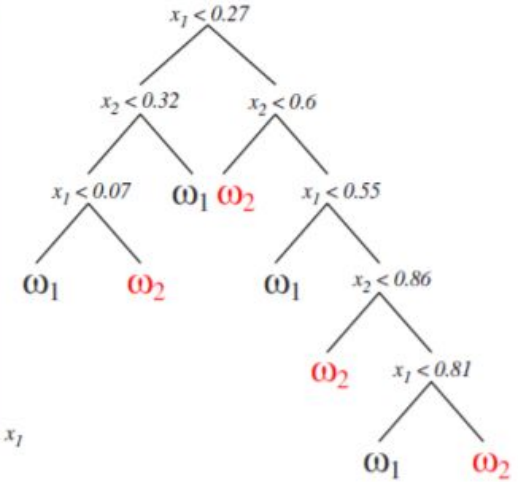
Weakness

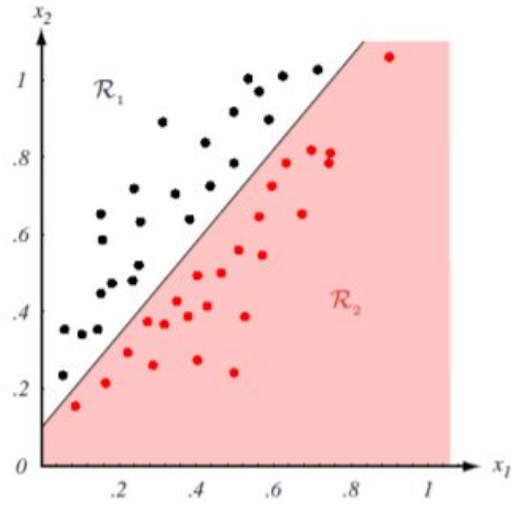
- ❏ Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
- ❏ Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.
- ❏ Decision tree can be computationally expensive to train. The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field must be sorted before its best split can be found. In some algorithms, combinations of fields are used and a search must be made for optimal combining weights. Pruning algorithms can also be expensive since many candidate sub-trees must be formed and compared.

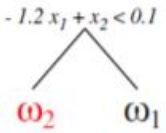
1. Classifiers

Decision Tree (Axis Alignment and Issues Associated)





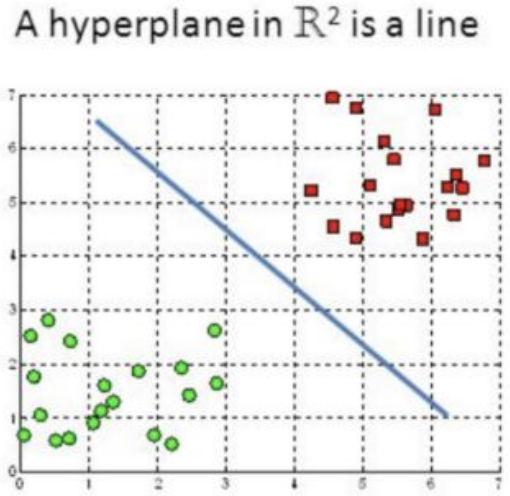


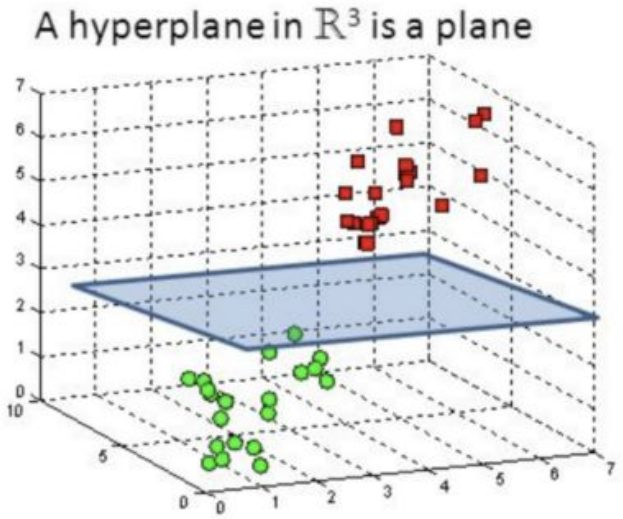


1. Classifiers

SVM

A support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection.
Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier.





1. Classifiers

SVM

In the SVM algorithm, we are looking to maximize the margin between the data points and the hyperplane. The loss function that helps maximize the margin is hinge loss. (More on this while Implementation)

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases}$$

1. Classifiers

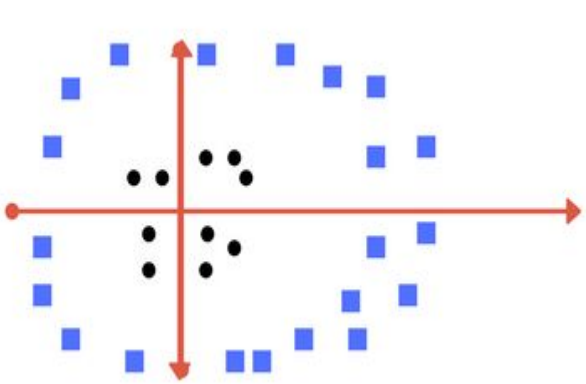
SVM (Tuning Parameters)

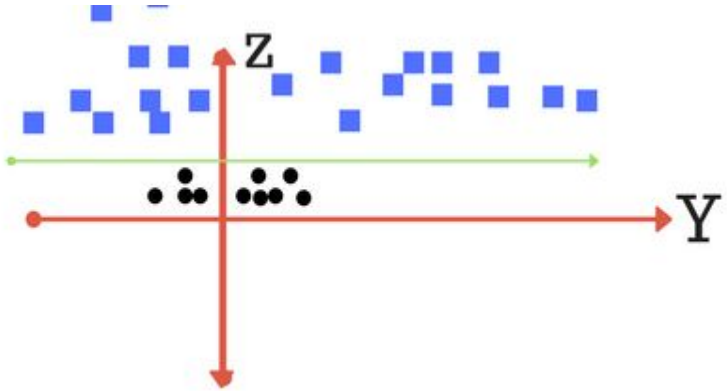
- Kernel
- Regularization
- Gamma and
- Margin

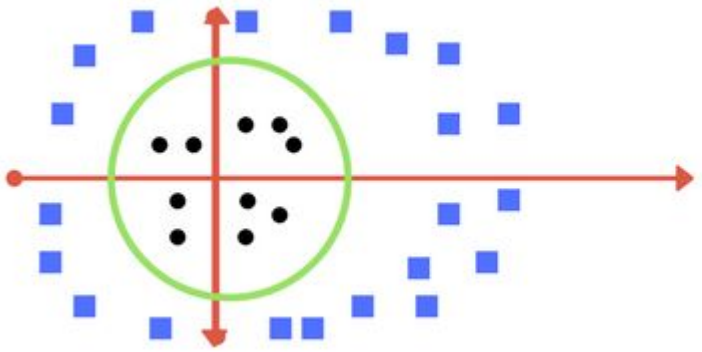
1. Classifiers

SVM (Kernels)

In the second one, the z values are $w = x^2 + y^2$







1. Classifiers

SVM (Kernels)

For linear kernel the equation for prediction for a new input using the dot product between the input (x) and each support vector (xi) is calculated as follows:
f(x) = B(0) + sum(ai * (x,xi))

The polynomial kernel can be written as K(x,xi) = 1 + sum(x * xi)^d and exponential as K(x,xi) = exp(-gamma * sum((x — xi^2)).

Kernel Trick: Polynomial and exponential kernels calculates separation line in higher dimension. This is called kernel trick

Read about Kernels from here: https://data-flair.training/blogs/svm-kernel-functions/

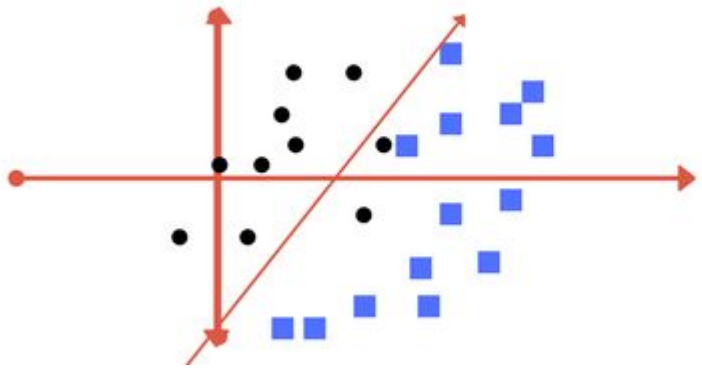
1. Classifiers

SVM (Regularizer)

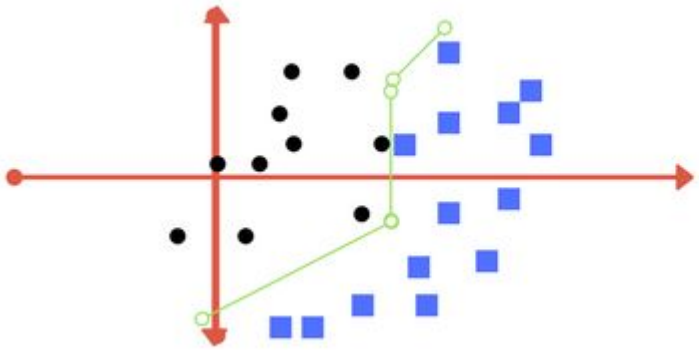
The C value, or regularization.

For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points.

Low C



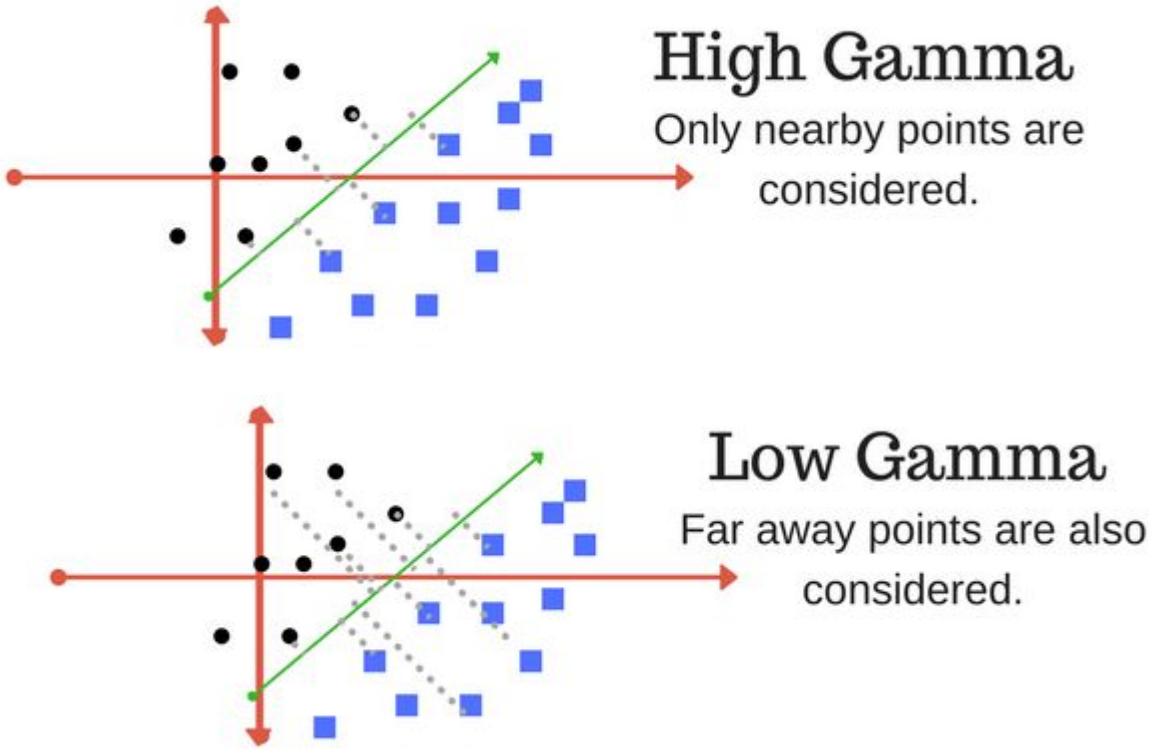
High C



1. Classifiers

SVM (Gamma)

The gamma parameter defines how far the influence of a single training example reaches, with low values meaning ‘far’ and high values meaning ‘close’. In other words, with low gamma, points far away from plausible seperation line are considered in calculation for the seperation line. Where as high gamma means the points close to plausible line are considered in calculation.



1. Classifiers

SVM (Margin)

A margin is a separation of line to the closest class points.