Naive Bayes – model is a generative model, so it estimates priors and conditional densities.

Linear regression is a supervised learning technique.

Logistic Regression – 2-class logistic regression is a linear classifier.
Logistic regression requires the use of gradient ascent.
Logistic regression the parameter n is called the learning rate. It control the speed at which changes happen to the w parameters.

Naïve bayes is generative classifier model learns a model of the point probability P(x,y)
logistic regression is Discriminative. Classifier models the posterior p(y | x) directly.

Discriminative model – P (y | x) take the form of a logistic sigmoid function – Call a logistic regression.
Logistic regression use the logistic function for modeling P (y | x), considering only the case y ∊ {0,1}.
The logistic function

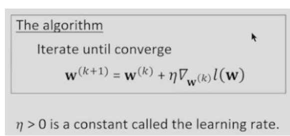
$$\sigma(t) = \frac{1}{1+e^{-t}} = \frac{e^t}{1+e^t}$$

P(y = 0| x) = 1 / 1 + e^ wtx = 1 - ∇ (w^tx)
P(y = 1| x) = e^wtx / 1 + e^wtx = ∇ (w^tx)
P(y=0|x) >= P(y=1|x)
1 >= e^wtx
0 >= wtx ⇔ if w^tx we classify x as 1
This is a linear classifier w^tx (linear function)
P(y|x) = ∇ (w^tx)^y (1 - ∇ (w^tx))^1-y

W^1 and W^2 two sets of parameters, whichever giving a larger P(y| x) should be a better parameter.

Call this L(w) the conditional likelihood.
The conditional log likelihood

We can use a commonly-used optimization technique, gradient descent/ascent, to find the solution. gradient ascent



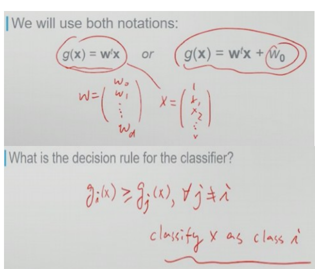
Linear Machine – Part 1 Basics
SVM is a time of linear machine.

Linear Classifier - $w^t x <= 0$, Class 0 , $w^t x > 0$, Class 1 - $g(x) = w^t x$ is called discriminant function. Is a linear classifier.
The learning task is to use the training samples to estimate the parameters of the classifier.

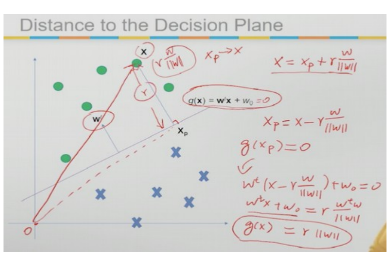
Linear discriminant functions give arise to linear decision boundaries.

If we can find at leat one vector w such that g(x) = w^tx classifies all samples.

We say the samples are linearly separable.
Solving for the Weight Vector
Theoretical: Lagrange or Karush-Kuhn-Tucker
In practice: eg. Gradient-descent-based search.



Margins for linear classifiers. The normal vector of the decision line/plane is w. g(x) = w^tx + w0, g(x) = 0



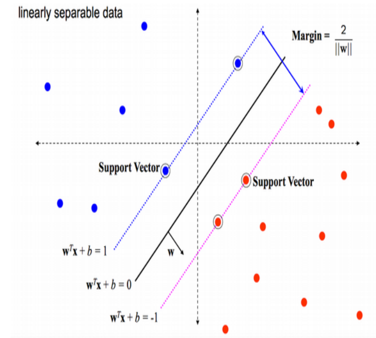
g(x) = 0 be a decision plane.

g(x) gives an algebraic measure of the distance from x to the decision plane.
r = g(x) / ||w||

for a given set of samples S, the margin is the smallest margin over all x in S.
For a given set, a classifier that gives rise the larger margin will be better.

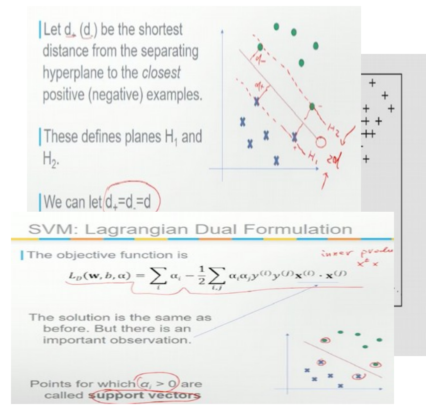
A nonlinear (quadralic) optimization problem with linear inequality constraints.
Reformulate the problem using lagrange multipliers α.
Lagrangian Primal or Dual Problem,

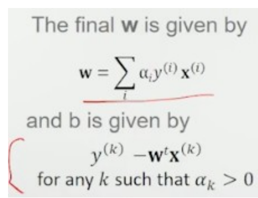
Lagrangian Primal



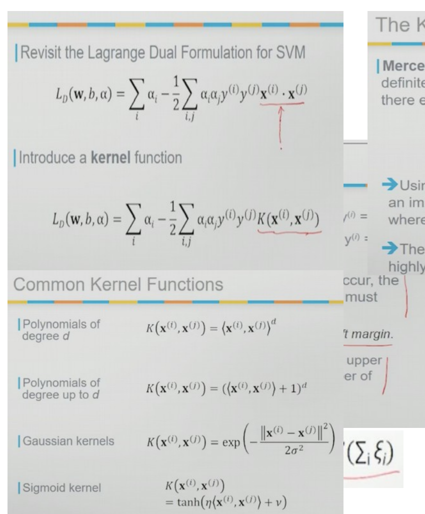
Support Vector Machine (SVM) – Discriminate Classifier formally define by a separating hyperplane.

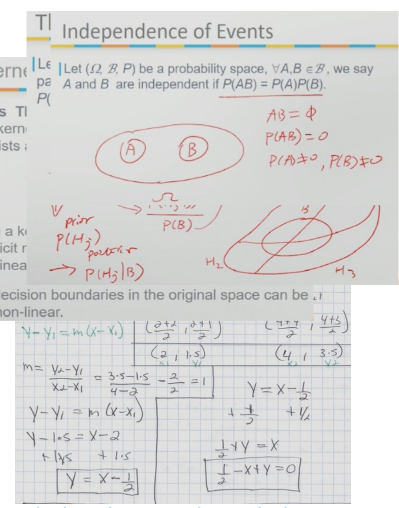
Support vectors are the data points that lie closest to the decision surface.
Goal of SVM, maximize the margin.





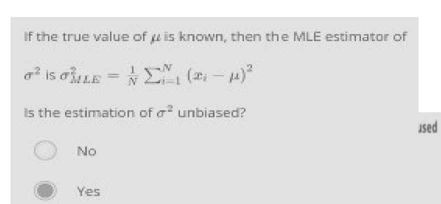
The region between two separate planes H1 : w^tx+b =1 H2: w^t +b = -1 as the margin the width is 2/ ||w||.
When the data has noise, there is a problem in drawing a clear hyperplan without misclassifying
Distance = SQRT ((x2-x1) ^2 + (y2 – y1)^2)
C is a parameter to control how much penalty is assigned to errors.





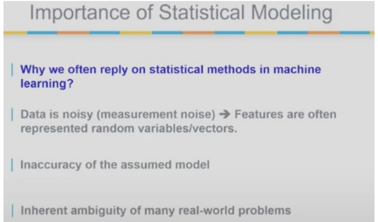
Maximize the margin and also minimize the error

For Large datasets, logistic perform better than naïve bayes.
Naïve Bayes Converges to its asymptotic estimates faster than logistic regression.
Generative classifier learns a model of the join probability p(x,y).
Discriminative classifier models the posterior p(y|x) directly.



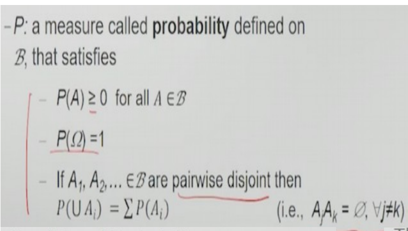
Logistic regression can be used as a classifier by using a threshold on the outcome of the logistic function and using the threshold to classify the inputs.

Preprocessing for Feature Extraction (segmentation, filtering, various transformation).
Good features should be invariant in some sense.
W1X1 + W2X2 + W0 = 0 <
Linear model.
Basic Machine Learning Paradigms.



**Supervised learning: the training samples have labels.
Regression and Classification
**Unsupervised learning: the training set is no labeled
Density estimation, Clustering
**Reinforcement learning: learning to take actions to maximize some notions of reward.
Categorical – Represent Characteristics
Ordinal–values or observations can be ranked. Rating scale
Numerical – Values or observations that can be measures

Trace Matrix = Sum mii = diagonal summation



Conditional Probability
P(B|H)=P(BH)/P(H)
=P(B∩H)/P(H)