UNIT - 3
STATISTICAL LEARNING

By Vishvajit Bakrola

MACHINE LEARNING

- Is a study and implementation of algorithms, in order to program computer to optimize a performance criterion using example data or past experience.
- Learning will become essential when,
  - ✓ Human expertise does not exist
  - ✓ Humans are unable to explain their expertise
  - ✓ Solution changes in time
  - ✓ Solution needs to be adapted to particular cases

MACHINE LEARNING - DEFINITION

- “Machine learning is a study of computer algorithms that allow computer programs to automatically improve through experience.”
- “The field of study that gives computers the ability to learn without being explicitly programmed.”
- “A computer program is said to learn from experience E with respect to some class of tasks T and the performance measure P, if its performance at task T, as measured by P, improves with experience E.”

DEFINING THE LEARNING TASK

- Improve on task T, with respect to performance metric P, based on experience E

T: Playing board game

P: Percentage of games won against an arbitrary opponent

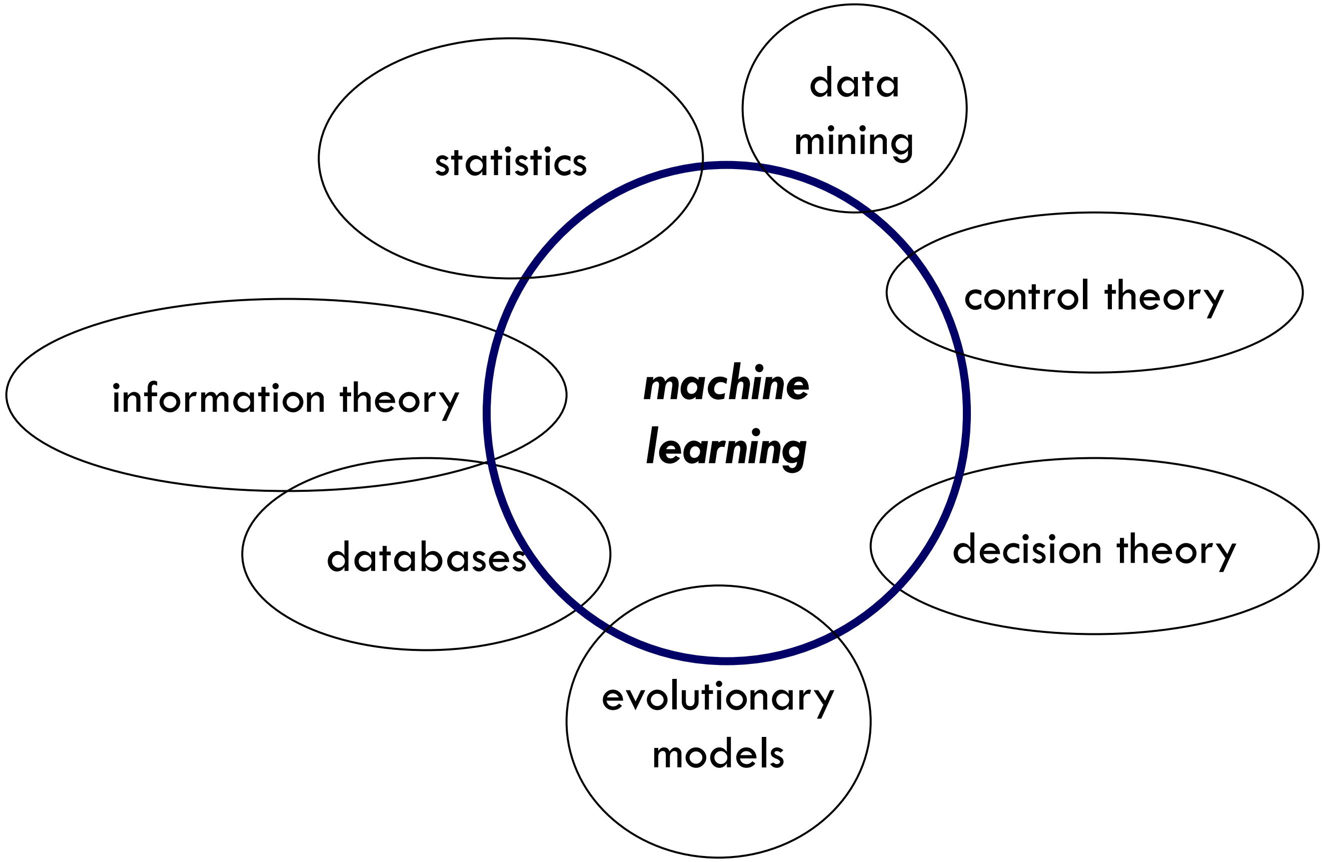
E: Playing practice games against itself

T: Recognizing hand-written characters

P: Percentage of characters correctly classified

E: Database of human-labeled images of handwritten characters

RELATED FIELDS



statistics

data mining

control theory

information theory

machine learning

databases

decision theory

evolutionary models

LEARNING ASSOCIATION

Basket Analysis – Finding association between products bought by customers.

- *If there is a customer who buy X typically also buy Y, and if there is a customer who buys X and does not buy Y, he or she is a potential Y customer.*
- This can be helpful for cross-selling.

Association Rule – Learning a conditional probability of the form P(Y|X), *Y is the product we would like to conditioned on X, which is a product which we know that the customer has already purchased.*

P(Biscuits|Tea) = 0.85 → We can define the rule that "85 percent of customers who buy **Tea** also buys **Biscuits**"

This can possibly P (Y|X, D), D:Customer Attributes

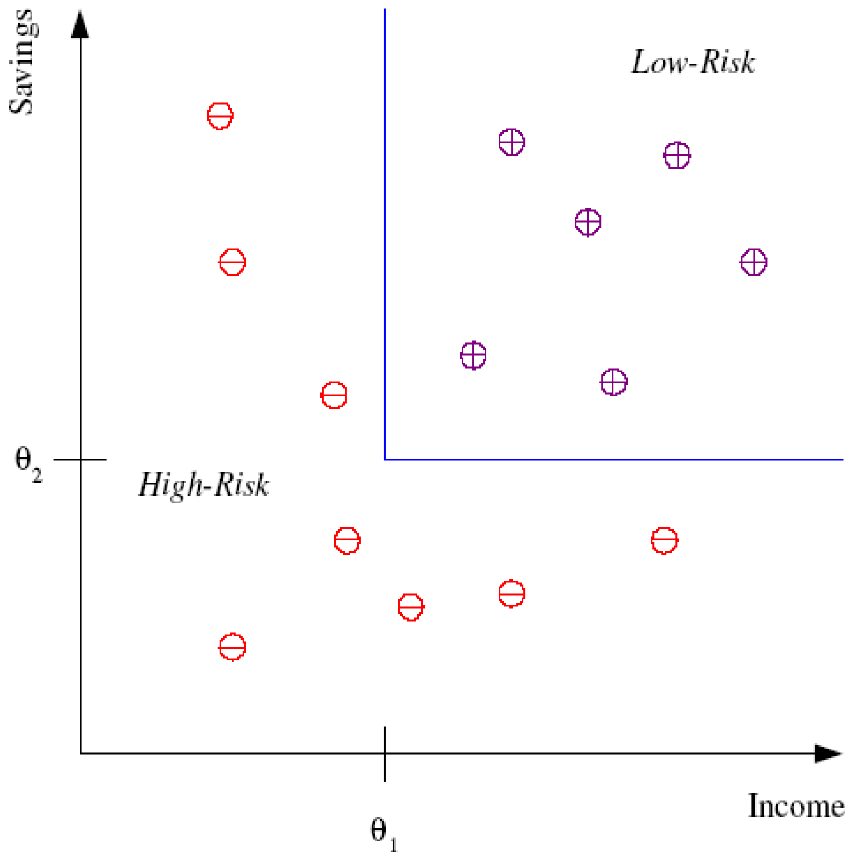
TYPES OF LEARNING

1. Supervised Learning
   - Training data + Desired outputs(labels) are given
2. Unsupervised Learning
   - Training data given without desired outputs
3. Semi-supervised Learning
   - Training data + Partial desired outputs available
4. Reinforcement Learning
   - Only collects rewards from sequence of actions

CLASSIFICATION

- When we need to assign individual unique label to classes/groups.
- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their *income* and *savings*.

Discriminant: IF *income* > θ_1 AND *savings* > θ_2 THEN low-risk ELSE high-risk



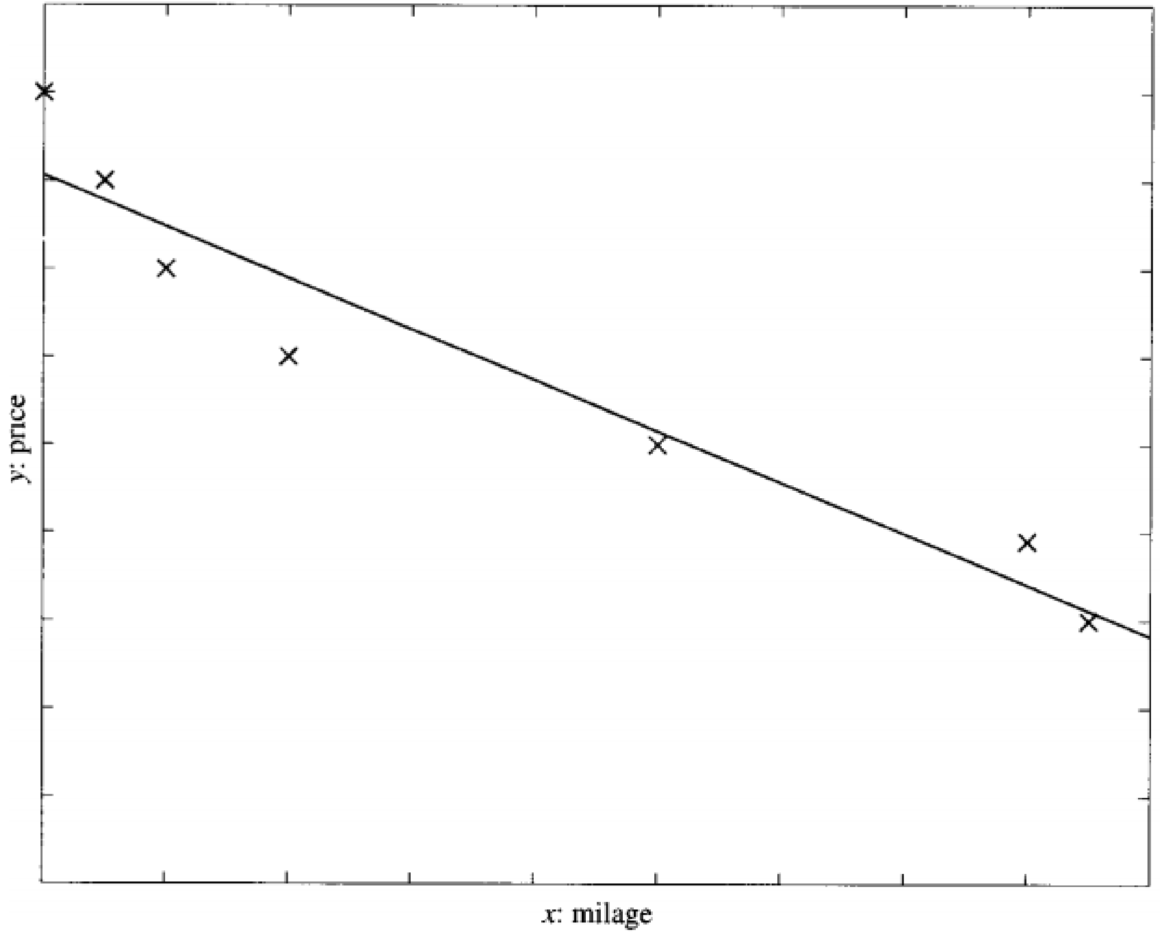
APPLICATIONS OF CLASSIFICATION

- ❑Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
- ❑Hand-written character recognition: Different handwriting styles
- ❑Medical diagnosis: From symptoms to illnesses
- ❑Others: E-mail spam, Image segmentation, Speech recognition, Genetic sequencing classification, etc.

REGRESSION

- Belongs to prediction types of problems, and their output is a number.
- Let, X is some attributes of a car and Y be the price of a car.
- Surveying the previous such transaction we can collect the training data. Fitting appropriate machine learning model gives us predicted price of car given the attributes.
- We use linear model *(simple equation of linear line i.e. $Y=mx+c$)*
- We get, $\mathbf{Y = WX + W0}$

CONTINUE...



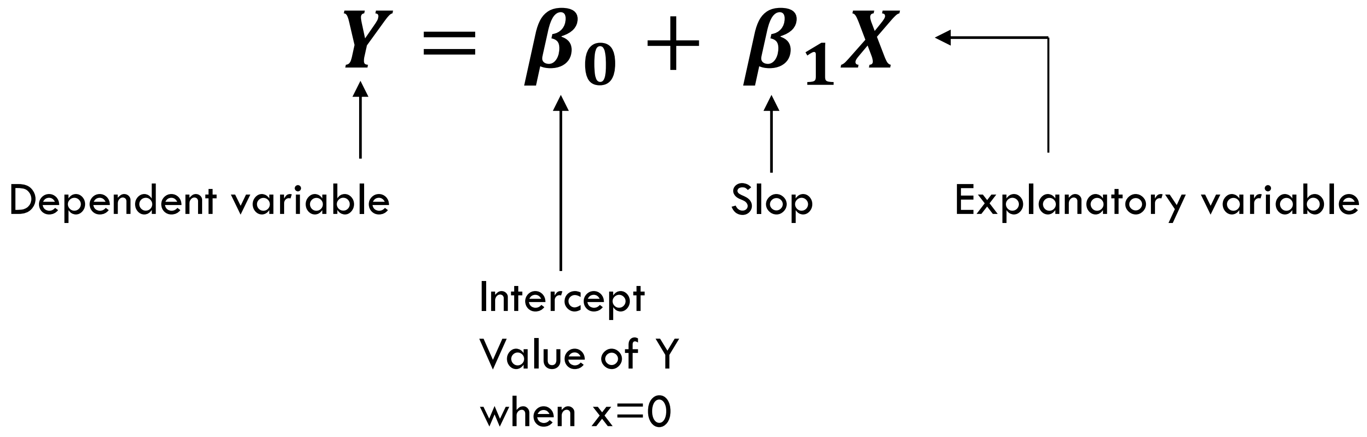
CONTINUE…

- The model is linear as W and W0 are the parameters optimized for best fit to the training data.
- If our data is type of non-linear, we can make out model quadratic or higher order polynomial.

$$\mathbf{Y} = \mathbf{W_2 X^2} + \mathbf{W_1 X} + \mathbf{W_0}$$

SIMPLE LINEAR REGRESSION

- We can predict score of one variable from the score of second variable.
- The variable(s) that we are predicting is **criterion variable** and we will refer it as Y.
- The variable we are basing upon for making predictions in called **predictor variable.**
- The general equation we get is,



$$Y = \beta_0 + \beta_1 X$$

Dependent variable

Intercept
Value of Y
when x=0

Slop

Explanatory variable

CONTINUE…

- Linear regression is an approach for modelling the relationship between a scalar dependent variable and one or more explanatory variables.
- The specific case of one explanatory variable is called **Simple Linear Regression.**
- NOTE:
  - ✓ When we use X to predict Y, called **Relationship estimation**
  - ✓ To estimate effect of X on Y, called **Forecast**

EXAMPLE

Student ID	xi	yi	(xi-X)	(yi-Y)	(xi−X) ^{2}	(yi−Y) ^{2}
1	98	85	20	8	400	64
2	85	95	7	18	49	324
3	80	70	2	-7	4	49
4	70	65	-8	-12	64	144
5	60	70	-18	-7	324	49
Sum	390	385				
Mean	**78**	**77**				

CONTINUE...

$b_1 = \Sigma [(x_i - x)(y_i - y)] / \Sigma [(x_i - x)^2]$

So, b1 = 494/730

Finally **b1 = 0.644**

CONTINUE...

- Using regression coefficient b1, we can solve for regression slope b0.

$$b_0 = y - b_1 * \mathbf{X}$$

$$= 77 - (0.644) * 78$$

$$= \mathbf{26.768}$$

CONTINUE...

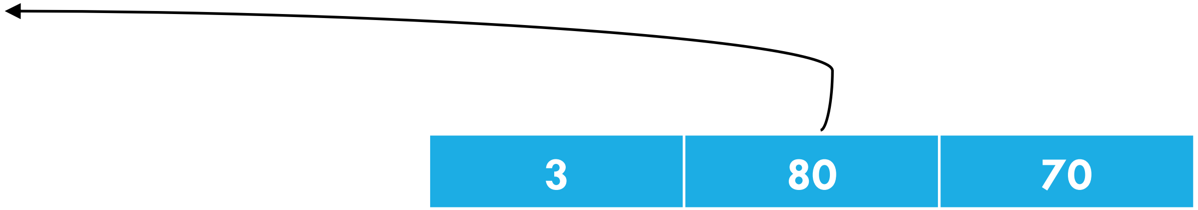
Now we have, **b1 = 0.644** and **b0 = 26.768**

We get,

$$Y = 26.768 + 0.644X$$

$$Y = 26.768 + 0.644*80$$

$$Y = 78.288$$



3	80	70

This is called **extrapolation.**

CONTINUE...

- Whenever, we use the regression equation we need to understand how well the equation fits the data.
- One solution to achieve these is to find **coefficient of determination.**

$$R^2 = \{\left(\frac{1}{N}\right) * \sum \frac{[(x_i - X) * (y_i - Y)]}{(\sigma_x * \sigma_y)}\}^2$$

N is Number of observations used to fit the model

$\boldsymbol{\sigma_x}$ is Standard deviation of x

$\boldsymbol{\sigma_y}$ is Standard deviation of y

CONTINUE...

$$R^2 = \{\left(\frac{1}{N}\right) * \sum \frac{[(x_i - X) * (y_i - Y)]}{(\sigma_x * \sigma_y)}\}^2$$

$$R^2 = \{\left(\frac{1}{5}\right) * \frac{470}{(12.083 * 11.255)}\}^2$$

$$R^2 = (\frac{94}{135.632})^2$$

$$R^2 = (0.693)^2$$

$$\boldsymbol{R^2 = 0.48}$$

Indicates 48% of variation in statistics grades (DV) can be explained by IV

CLASSIFICATION

- The understanding of boundary conditions that can be use to determine each target class in training data.
- Once the boundary conditions are determined the next task is to predict the target class.
- The process formally known as **classification.**
- Examples:
  - ✓Analysis of student data, whether he/she will buy a laptop or not. (**Target class: Yes or No**)
  - ✓Classifying fruits using associated features like color, shape, size, weight, etc. **(Target class: Name of the Fruit)**
  - ✓Student classification using features of student-uniform. (**Target class: Name of institute**)

CONTINUE…

- **Classifier:** A technique or an algorithm that maps input data to specific category.
- **Classification model:** Retrieve some meaningful conclusion from the input data given during training in order to predict the class labels.
- **Feature:** Is an individual and unique measurable property being observed from the data.
- **Binary classification:** The task of classification having possible outcomes.
- **Multi-class classification:** The task of classification having more than two outcomes.
- **Multi-label classification:** The task of classification having more than two class labels. One or more class labels may be predicted for each example.

GENERAL TYPES OF CLASSIFICATION ALGORITHMS

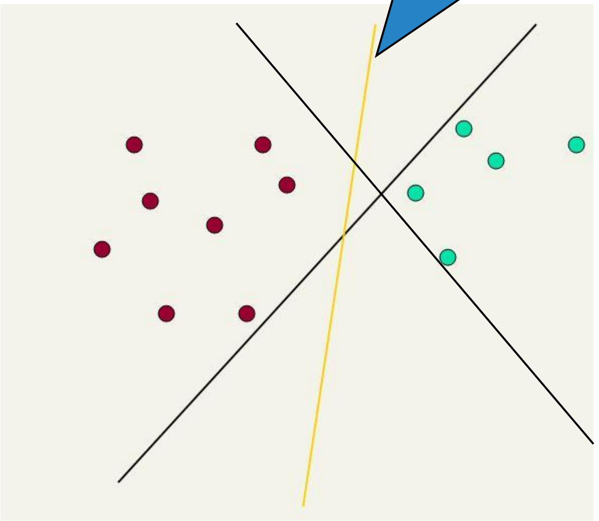
- Linear classification
  - Logistic Regression
  - Naïve Bayes Classifier
  - Linear discriminant
- SVMs
- Quadratic classifiers
- Decision Trees
- Neural Networks
- LVQs

SUPPORT VECTOR MACHINES

- Linear classifier is there but which hyperplane to select?
- Lots of possible solutions for *a*, *b*, *c*.
- Some methods find a separating hyperplane, but not the optimal one.

Support Vector Machine (SVM) finds an optimal solution.

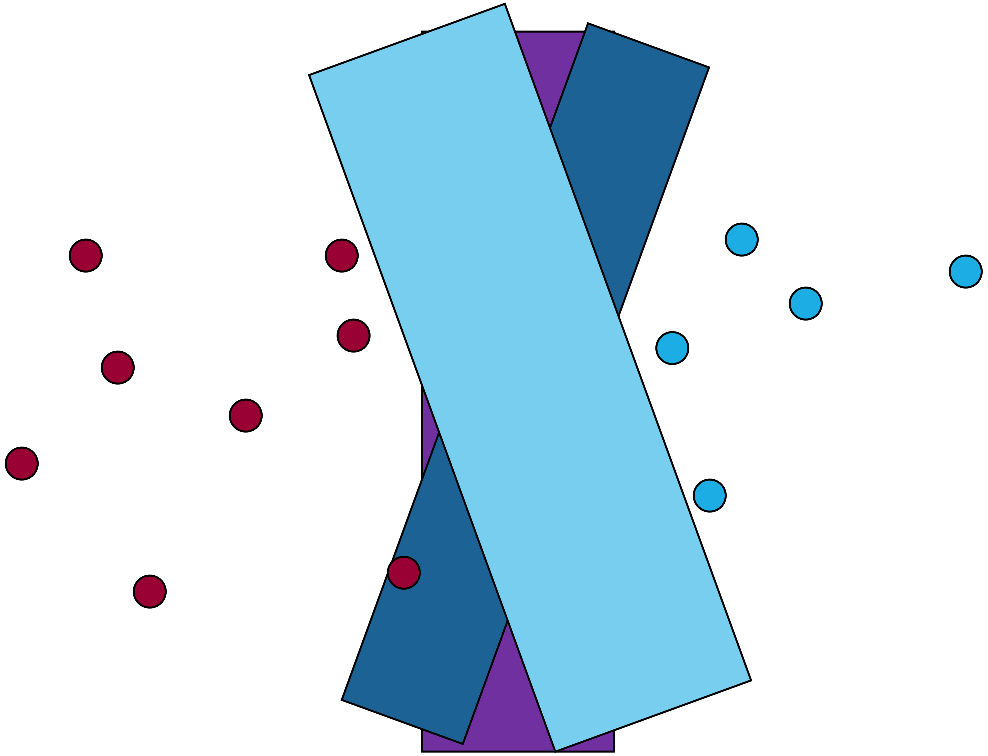
  - Maximizes the distance between the hyperplane and the “difficult points” close to decision boundary
  - One intuition: if there are no points near the decision surface, then there are no uncertain classification decisions.



This line represents the decision boundary:
$ax + by - c = 0$

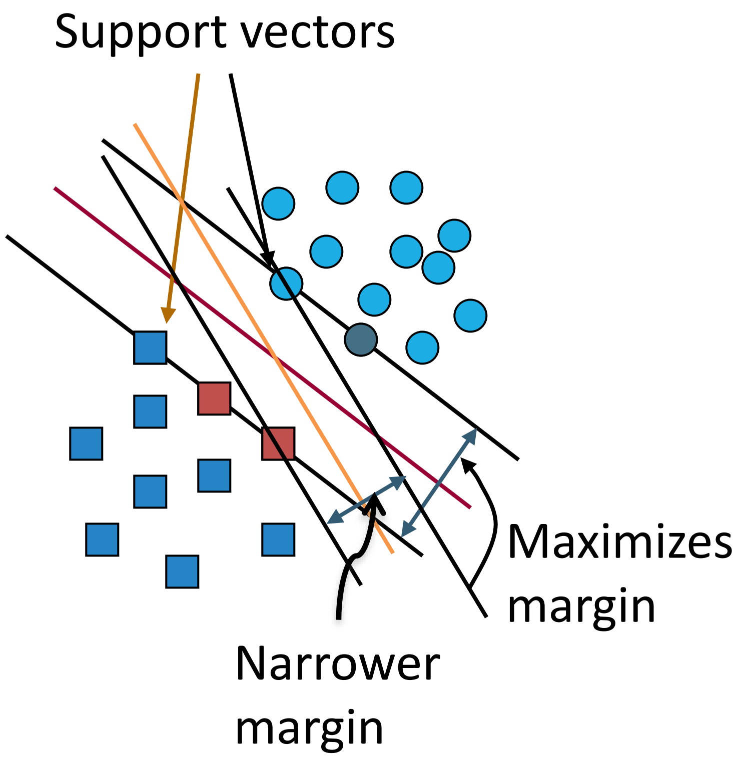
CONTINUE...

- If we have to place a fat separator between classes, we have less choices, and so the capacity of the model has been decreased.



CONTINUE…

- SVMs **maximize the margin** around the separating hyperplane.
- The decision function is fully specified by a subset of training samples, the **support vectors.**
- Solving SVMs is a **quadratic programming problem.**
- Seen by many as the most successful text classification method.



CONTINUE...

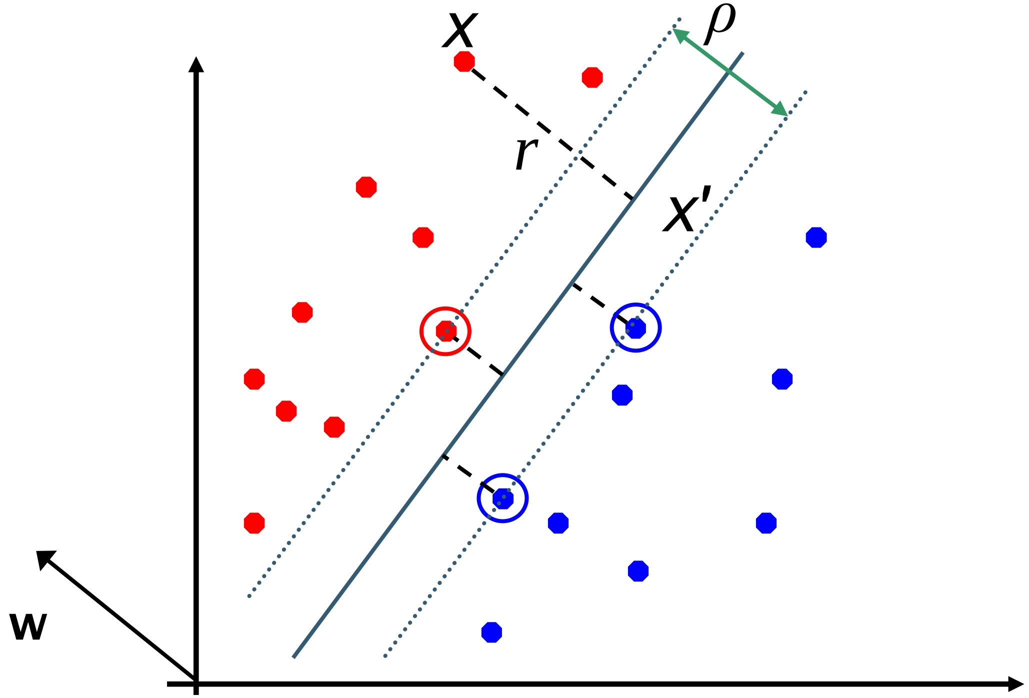
- **w**: decision hyperplane normal vector
- $\mathbf{x}_i$: data point *i*
- y_i: class of data point *i* (+1 or -1)
- Classifier is: $\mathbf{f(x_i) = w^T x_i + b}$
- Functional margin of $\mathbf{x}_i$ is: $y_i(\mathbf{w}^T\mathbf{x}_i + b)$
  - But note that we can increase this margin simply by scaling **w, b**....
- Functional margin of dataset is twice the minimum functional margin for any point
  - The factor of 2 comes from measuring the whole width of the margin

CONTINUE...

- Distance from example to the separator is

$$r = y\frac{\mathbf{w}^T\mathbf{x}+b}{\|\mathbf{w}\|}$$

- Examples closest to the hyperplane are ***support vectors.***
- ***Margin*** ρ of the separator is the width of separation between support vectors of classes.
- Dotted line **x′−x** is parallel to **decision boundary.**



CONTINUE...

- Following two constraints follow for a training set $\{(\mathbf{x_i}, y_i)\}$.

$$\mathbf{w^T x_i} + b \geq 1 \quad \text{if } y_i = 1$$

$$\mathbf{w^T x_i} + b \leq -1 \quad \text{if } y_i = -1$$

- Then, since each example's distance from the hyperplane is $r = y\frac{\mathbf{w}^T\mathbf{x}+b}{\|\mathbf{w}\|}$
- The margin is: $r = \frac{2}{\|\mathbf{w}\|}$

DISCRIMINANT FUNCTIONS

- A discriminant is a function that takes an input vector x and assigns it to one of k-classes, denoted by C_k.
- K>2 is for multiclass problem.
- The simplest representation of a linear discriminant function is obtained by taking a linear function of input vector so that,

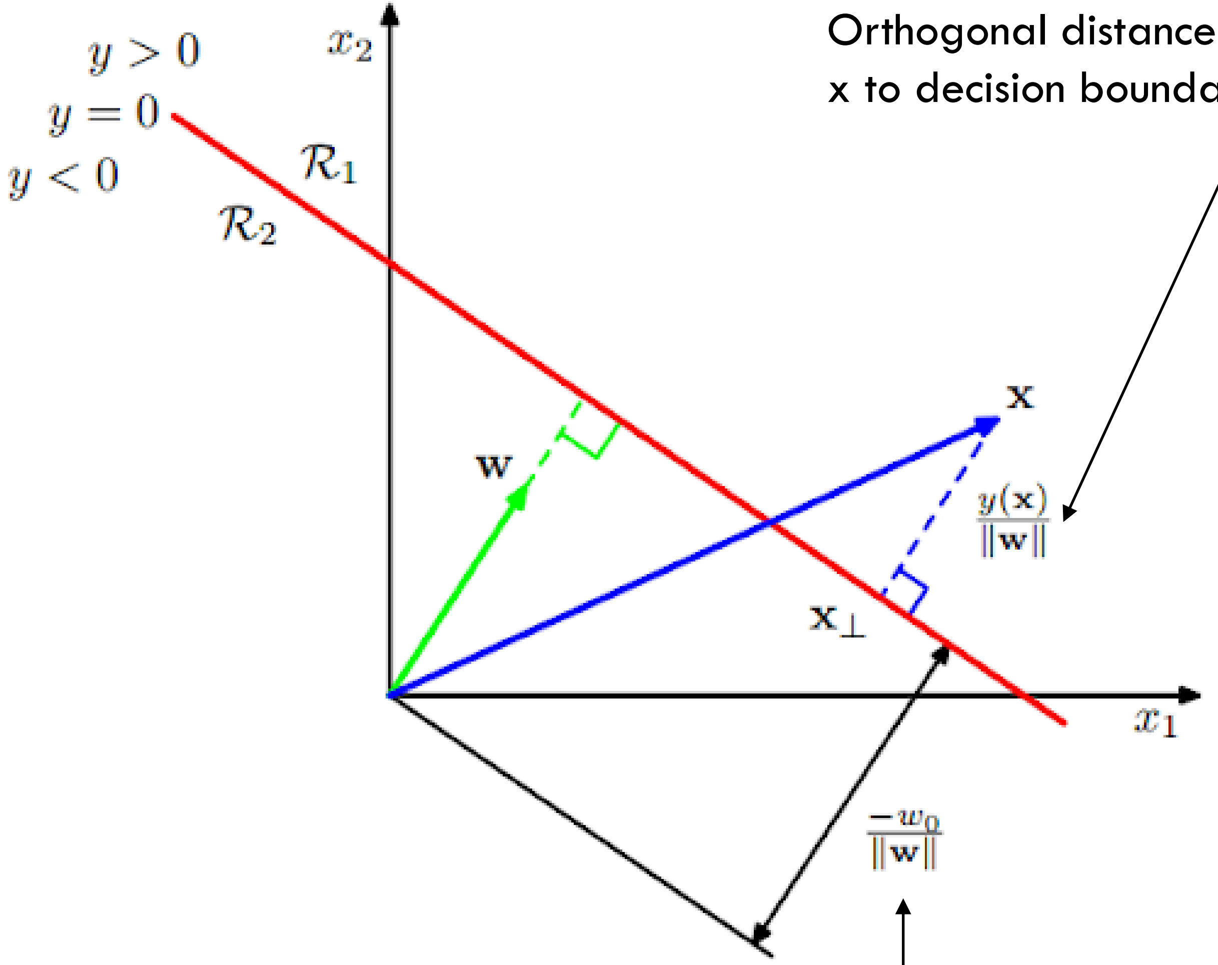
$$\boldsymbol{y(x)} = \underset{\substack{\uparrow \\ \text{Weight Vector}}}{\boldsymbol{W^T}} \cdot \boldsymbol{X} + \underset{\substack{\uparrow \\ \text{Bias}}}{\boldsymbol{W_0}}$$

CONTINUE…

- An input vector x is assigned to class,

$\boldsymbol{C_1 = \ y(x) \geq 0}$ **and** $\boldsymbol{C_2 = otherwise}$

- The corresponding decision boundary is therefore defined as $y(x) = 0$.
- $y(x) = 0$ corresponds to a (D-1) dimensional hyperplane within the D-dimensional input space.

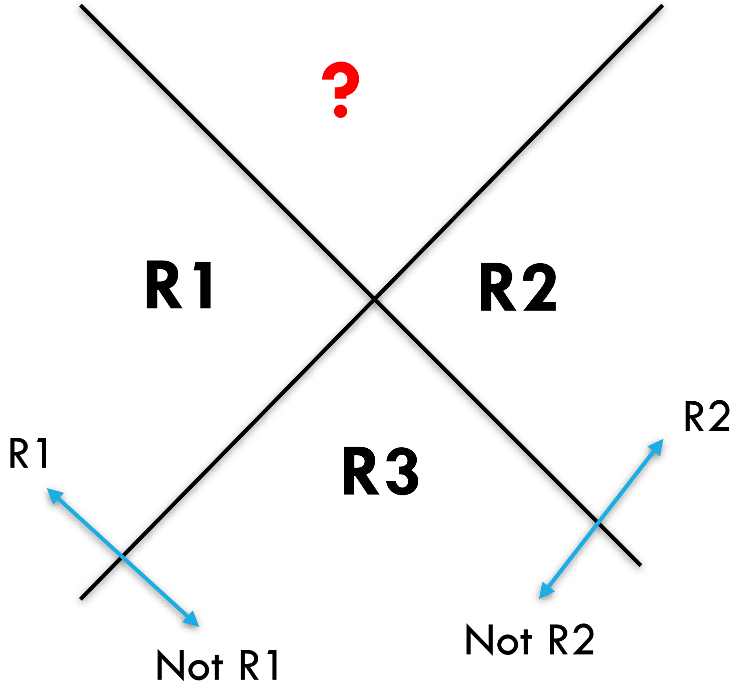


Orthogonal distance from x to decision boundary

Displacement from the Origin

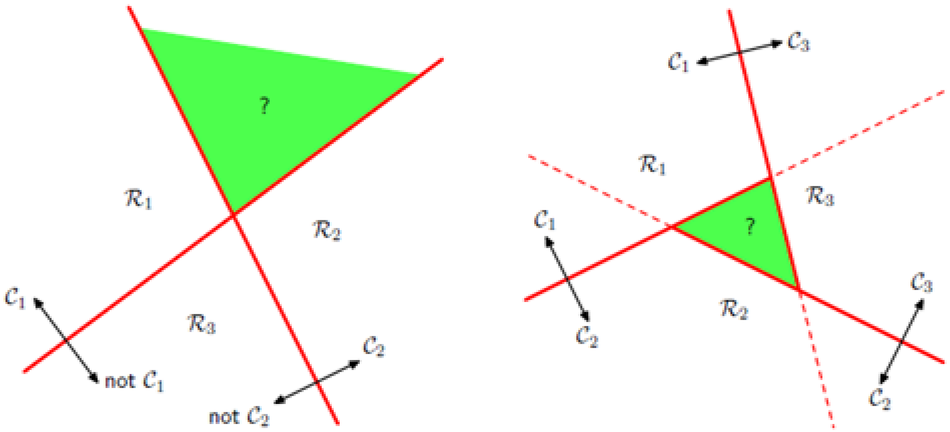
CONTINUE...

- It is extension of linear discriminant to k>2 classes.
- Use of k-1 classifiers where, each solves a two class problem of separating points in a particular class C_k from points not in that class, known as a **"one-versus-the-rest" classifier.**



CONTINUE…

- The examples involving three classes, where this approach leads to regions of input space that are ambiguously classified.
- An alternative is to introduce k(k-1)/2 binary discriminant function. One fore every possible pair of classes. This is known, **“one-versus-the-one” classifier**.



VARIANCE AND COVARIANCE

- Variance and Covariance are a measure of the “spread” of a set of points around their center of mass (mean).
- Variance – measure of the deviation from the mean for points in one dimension.
- Covariance as a measure of how much each of the dimensions vary from the mean with respect to each other.
- Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions e.g. number of hours studied & marks obtained.
- **The covariance between one dimension and itself is the variance**
- Used to find relationship between dimensions in the high dimensional data sets, where visualization is difficult.

CONTINUE...

- Covariance between x and y,

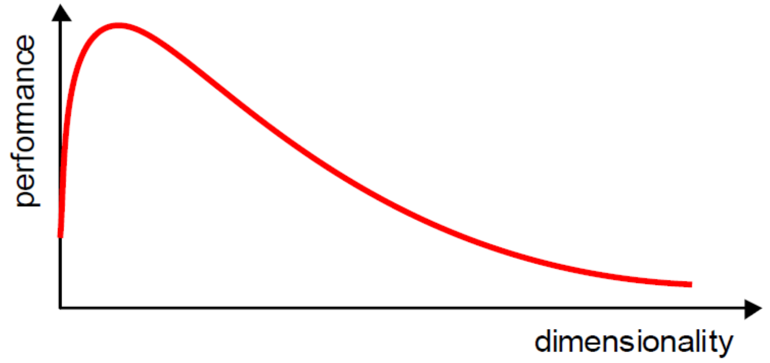
$$(x, y) = \frac{\sum_{i=1}^{n}(\bar{X}i\ - X)(\bar{Y}i\ - Y)}{(n-1)}$$

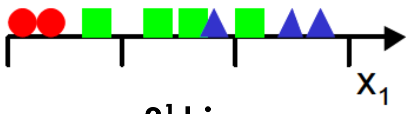
CONTINUE…

- If we have a 3-dimensional data set (x, y, z), then we first measures the covariance between the x and y dimensions, then y and z dimensions, and then x and z dimensions.
- A **positive value** of covariance indicates **both dimensions increase or decrease together** e.g. as the number of hours studied increases, the marks in that subject increase.
- If **covariance is zero**: the **two dimensions are independent** of each other e.g. heights of students vs the marks obtained in a subject
- A **negative value** indicates while **one increases the other decreases, or vice-versa.**

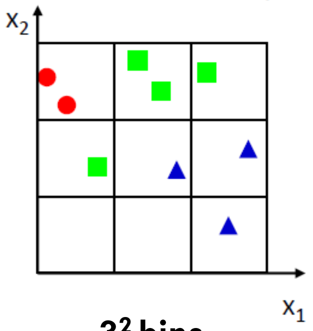
DIMENSIONALITY REDUCTION

- Increasing the number of features will not always improve classification accuracy – **Curse of Dimensionality.**
- In practice, the inclusion of more features might actually lead to worse performance.
- The number of training examples required increases exponentially with dimensionality **d** (i.e., k^d).

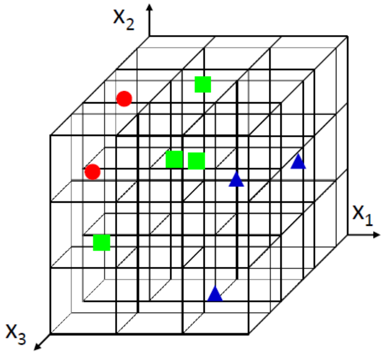




3^1 bins



3^2 bins



3^3 bins

CONTINUE…

- **Visualization** – Projections of higher dimensional data to 2D or 3D.
- **Removal of noise** – Removing noise gives the clarity in data and positive impacts on accuracy.
- **Compression of data** – Ultimately leads to efficient storage and easy data retrieval.

CONTINUE…

- Our motto is to choose an optimum set of features of lower dimensionality to improve classification accuracy.
- **Feature extraction:** To finds a set of new features (i.e., through some mapping f()) from the existing features.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ . \\ . \\ x_N \end{bmatrix} \xrightarrow{f(\mathbf{x})} \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_K \end{bmatrix}$$

The mapping f()
could be linear or
non-linear.
Here, K<N

CONTINUE...

- **Feature selection:** chooses a subset of the original features.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ x_N \end{bmatrix} \to \mathbf{y} = \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \cdot \\ \cdot \\ x_{i_K} \end{bmatrix}$$

- Linear combinations are particularly attractive because they are simpler to compute and analytically tractable.
- Commonly used linear feature extraction methods:
  - Principal Components Analysis (PCA): Seeks a projection that **preserves** as much **information** in the data as possible.
  - Linear Discriminant Analysis (LDA): Seeks a projection that **best discriminates** the data.
  - Few other methods are - Projection Pursuit, Independent Component Analysis or ICA, Isomap, Locally Linear Embedding or LLE, etc.

LINEAR DISCRIMINANT ANALYSIS

- Is dimensionality reduction technique used as a pre-processing.
- As we know that the main goal of dimensionality reduction is to remove redundant and dependent features by transforming them in lower dimensions.
- LDA is supervised technique, as it takes labels into consideration.
- In LDA, we first calculate the separability between two classes and then the distance between mean and sample of each class.

CONTINUE...

- Step – 1: To calculate distance between mean of different classes.

$$S_b = \sum_{i=1}^{g} N_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T$$

CONTINUE…

- Step – 2: To calculate distance between mean and sample of each particular class.

$$S_w = \sum_{i=1}^{g} (N_i - 1) S_i = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (x_{i,j} - \bar{x}_i)(x_{i,j} - \bar{x}_i)^T$$

CONTINUE…

Step – 3: To project the lower dimensional space which maximizes the variance between classes and minimizes the in-class variance.

$$P_{lda} = \arg\max_{P} \frac{\left|P^T S_b P\right|}{\left|P^T S_w P\right|}$$

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

- Statistic can be any value that is calculated from a given sample.
- In the process of inference with statistic, we make decision using the information provided by a sample.
- One of the approaches is parametric approach, where we assume that the sample is drawn from some distribution that follows our known model.
- It can be drawn from small number of parameters, and once those parameters are known the whole model is known to us.

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

- MLE is used for estimation of critical **parameters** of our model.
- In machine learning we are referring a model as a process that ultimately results us the data that are observed.
- I.e. object classification with object properties, spend on advertisement vs. revenue generated, etc.
- All such models is having its own set of parameters, that ultimately defines what exactly the model is.

CONTINUE…

- From linear equation $y=mx+c$, where x is the spending on advertisement and y will be revenue generated.
- m and c will be the parameters for this model. We understand the different values of this model gives different association between values of x and y.
- Ultimately parameters are defining how model will react on the data.
- MLE is a method that determines the values of such parameters of our model.
- Here, MLE returns parameters values that maximize the likelihood of our sample.

CONTINUE…

- **Bernoulli distribution** – There are two outcomes that whether the event will occur or not. i.e. **any given instance is a positive example of a class or not.**
- **Multinomial distribution** – It is generalization of Bernoulli distribution, where instead of two states the outcome of a random event is one of K mutually exclusive states. Let x1, x2…xk are indicator variables where xi is 1 if the outcome is state i and 0 otherwise.
- **Gaussian distribution** – distributed with value of mean and variance.

CONTINUE…

- Assume, we are having **three data points** that has been generated using some data process. **Say, 9, 9.5 and 11**.
- Ultimately we are trying to calculate the total probability of observing all the data.
- Simply, the joint probability distribution of all the observed data points.
- The PDF for observing single data point x, using gaussian distribution:

$$P(\underset{\substack{\uparrow \\ \text{Data point}}}{x};\underset{\substack{\uparrow \\ \text{Parameters of model}}}{\mu,}\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

CONTINUE...

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$P(9, 9.5, 11;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9.5-\mu)^2}{2\sigma^2}\right) \\ \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(11-\mu)^2}{2\sigma^2}\right)$$

$$\ln(P(x;\mu,\sigma)) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(9-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(9.5-\mu)^2}{2\sigma^2} \\ + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(11-\mu)^2}{2\sigma^2}$$

$$\ln(P(x;\mu,\sigma)) = -3\ln(\sigma) - \frac{3}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\left[(9-\mu)^2 + (9.5-\mu)^2 + (11-\mu)^2\right]$$

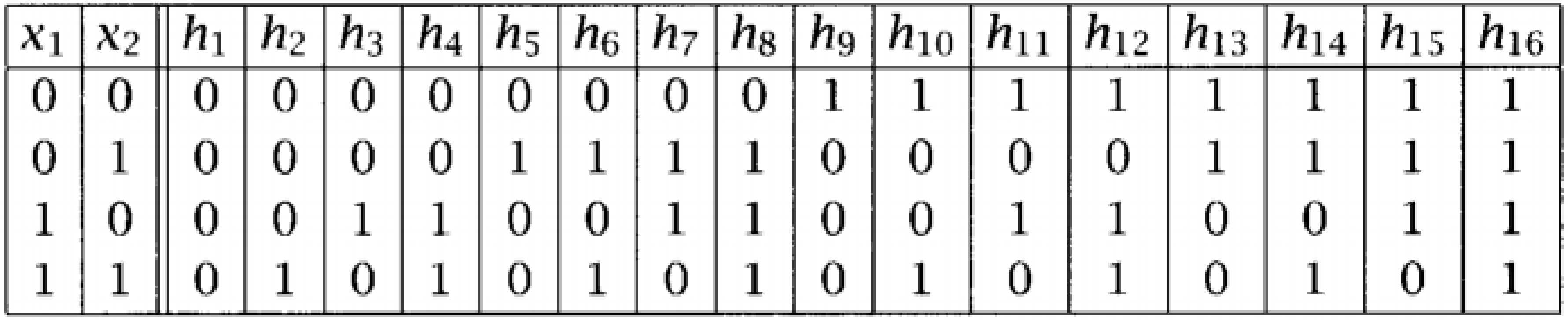
CONTINUE...

$$\frac{\partial \ln(P(x;\mu,\sigma))}{\partial \mu} = \frac{1}{\sigma^2}\left[9 + 9.5 + 11 - 3\mu\right].$$

$$\mu = \frac{9 + 9.5 + 11}{3} = 9.833$$

MODEL SELECTION AND GENERALIZATION

- To understand learning taking example of Boolean function. Where all the inputs and outputs are binary.
- There are 2^d possible ways to write d binary values and so with d inputs, the training set has at most 2^d examples.



x_1	x_2	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	h_{11}	h_{12}	h_{13}	h_{14}	h_{15}	h_{16}
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

With two inputs there are four possible cases and sixteen possible Boolean functions

CONTINUE...

- Here each training examples removes half of the hypotheses, specifically those whose guesses are wrong.
- For example, in the case of x1=0 and x2=1 the output is 0. This removes h5, h6, h7, h8, h13, h14, h15, h16.
- This is one of the ways to see and understand the learning. With more examples we removes those hypotheses that are not consistent with the training data.
- In our cases of Boolean function, to end us with single hypothesis we need to see all the 2^d training examples.
- If training set contains only a small subset of possible instances, means the known true output has very small set of case to justify itself, then the solution is not unique.

CONTINUE…

- After seeing all the N examples cases, there remain 2^2d-N possible functions.
- This is referred as **ill-posed problem – Where the data by itself are not sufficient to find a unique solution.**
- The same problem also exists in many learning applications, in classification and in regression as well.
- So the learning is ill-posed the data itself are not sufficient to find the solution, we required to make some additional assumptions to have a specific unique solution with the existing data only, that we are having.
- The set of assumptions we have to maek learning possible is called **inductive bias** of learning algorithm.

CONTINUE…

- Thus learning is not possible without inductive bias, and the process to choose right bias is called **model selection**.
- We need to understand that the scope of the machine learning is rarely to replicate the training data but the prediction for new cases.
- How well the model trained on the training set predicts the right outcome for new instances is called **generalization**.
- For the best generalization we should math the complexity of the hypothesis with the complexity of the function that underlying the data.

CONTINUE..

- If the hypothesis is less complex than the function, we have **underfitting**. As we try to fit a line on data sampled from third order polynomial.
- In such cases, as we increase the complexity both the training error and validation error decrease.
- But, if we have a hypothesis that is too complex, the data is not enough to constraint it and we may end up with a bad hypothesis.
- For example, fitting sixth order polynomial to noisy data sampled from a third order polynomial. This is called **overfitting**. In such case, having more training data helps but only up to a certain point.

CONTINUE…

➢**The triple trade-off** – All learning algorithms those are trained from example data, there is a trade-off between three factors:

1. The complexity of hypothesis we fit to data, means the capacity of the hypothesis class.
2. The amount of training data.
3. The generalization error on new examples.

CONTINUE…

- The amount of training data increases, the generalization error decreases.
- The complexity of the model increases, the generalization error decreases first and then start to increase.
- The generalization error of a complex hypothesis can be kept in check by increasing the amount of training data but only up to a point.
- We can measure the generalization ability of a hypothesis, namely the quality of its inductive bias, if we have access to data outside the training set.
- We perform this by dividing the training set into two parts, we use one part for training and other called validation set, which is used to test generalization ability.

CONTINUE…

- Considering the large enough training and validation set, the hypothesis that is the most accurate on the validation set is considered as best one – the one that has the best inductive bias. The process is called cross-validation.
- In the case, when we need to report the error to give an idea about the expected error of out best model, we should not use the validation error.
- As we have used the validation set to choose the best model, and it has effectively become a part of training set.
- We need a third set, which should not be a part of training or validation set – called test set.

EVALUATING AN ESTIMATOR WITH BIAS AND VARIANCE

- Let X be a sample from a population specified up to a parameter θ and let d=d(x) be an estimator of θ.
- To evaluate, the quality of this estimator, we can measure how much it is different from θ, that is $(d(x) - \theta)^2$. As it is a random variable it depends of the samples, we need to average it over possible X and consider r(d, θ) and so the mean square error of an estimator d defined as:

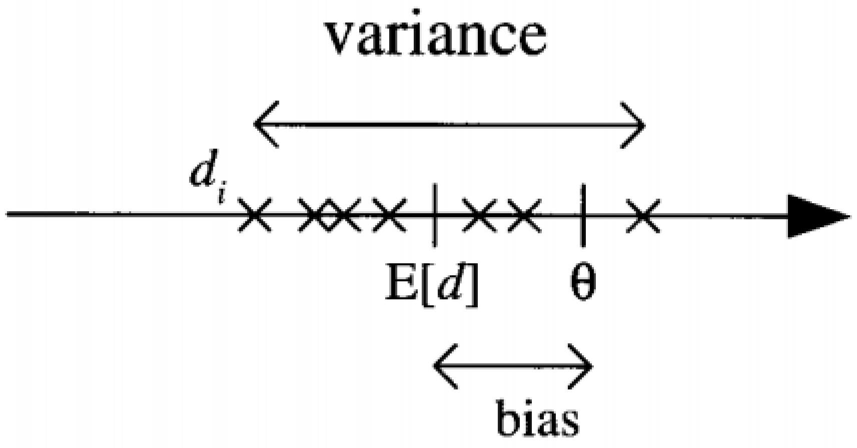
r(d, θ) = E[$(d(x) - \theta)^2$]

The bias of an estimator is given as,

$b_\theta(d) = E[d(x)] - \theta$

CONTINUE…

- θ is the parameter to be estimated.
- di are several estimates over different samples.
- Bias is the difference between the expected value of d and θ.
- Variance is how much di are scattered around the expected value.
- **We would like both to be small.**



REFERENCES

[1] Introduction to Machine Learning – Ethem Alpaydin, MIT Press