Session 3

- Case Studies – 15 mins
- Recap of Session 2
- Basic theory of Machine Learning
- Walkthrough of some supervised learning models

Session 3

- Case Studies – 15 mins
- Recap of Session 2
- Basic theory of Machine Learning
- Walkthrough of some supervised learning models

Recap

- Numpy
  - Similar to python list but lot more powerful
  - Used for data preparation before you feed the same to algos
  - Be careful of the dimensions – broadcasting works but not always
    - Myarray[None, :] – will add a new axis to data in front
- Chatbots using Dialogflow and Slack
  - Most critical is to model well the interaction you want to have
  - Uses Deeplearning driven NLP techniques to allow fuzziness in your dialogs

Recap

- Linear Algebra
  - Vectors, Matrices
  - Dot product – gives something called "Cosine similarity"
    - one of the ways to find similarity between two vectors
  - Most imp thing to remember:
    - get the dimensions of your data properly aligned

Additional numpy exercises

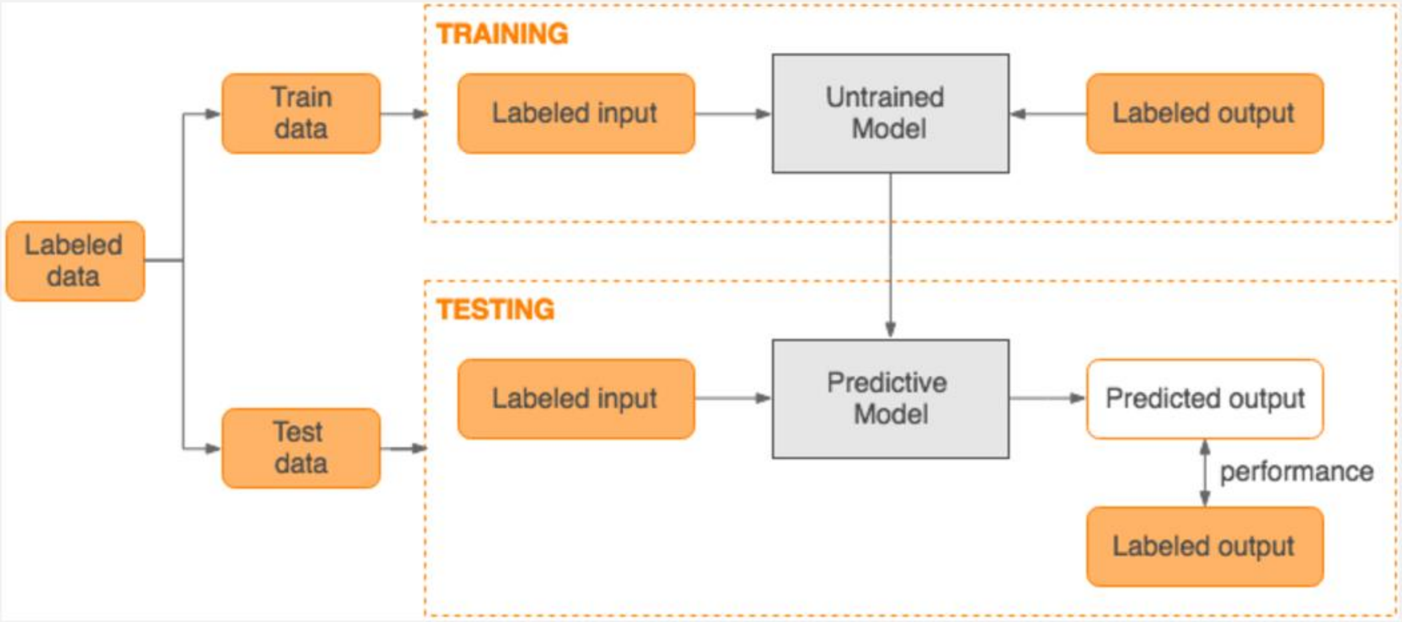
- Exercises
  - http://www.scipy-lectures.org/intro/numpy/exercises.html
    - you can try, a little harder but fun
  - http://www.labri.fr/perso/nrougier/teaching/numpy.100/
    - Seems interesting. Have not tried it personally

Session 3

- Case Studies – 15 mins
- Recap of Session 2
- Basic theory of Machine Learning
- Walkthrough of some supervised learning models

What is ML – Step 1

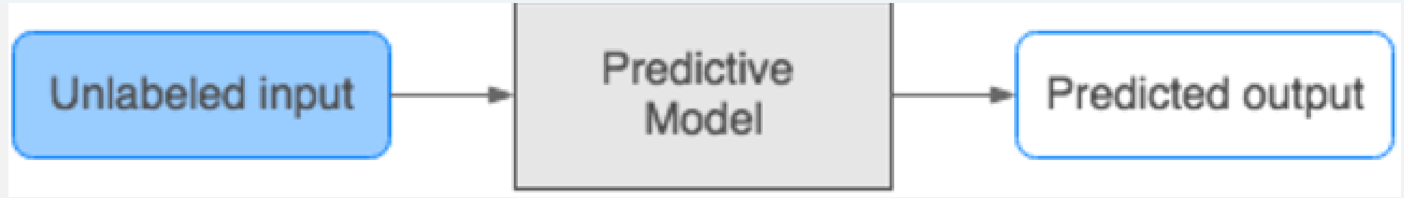
- We take a sample of data and split it into training and testing set
- Use training data to build predictive model
- Use testing data to check the quality of model



Splitting labeled data into training and testing sets

What is ML – Step 2

- Use predictive model to future data

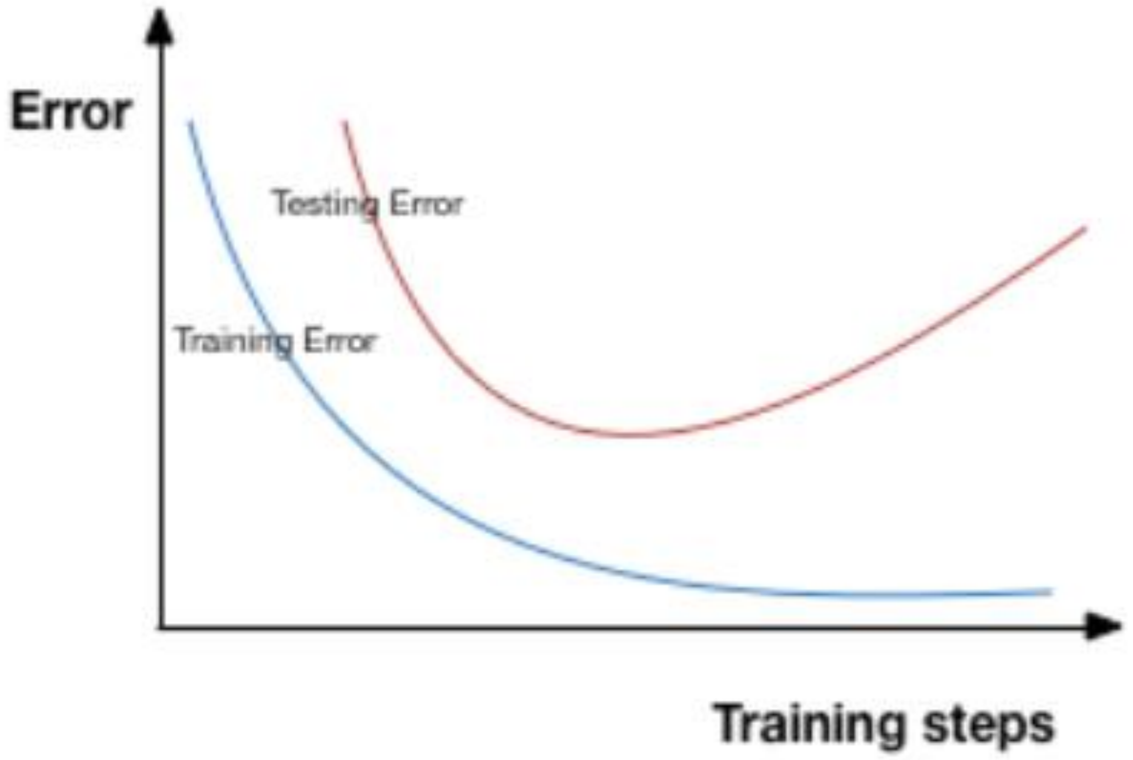


- We do not see complete data during training – only a sample
  - We want to have a way so that "training error" stays close to "testing error"
  - And "testing error" close to "actual error" of the model if it was evaluated on complete data

Theory provided by Probability

Please refer Appendix A for a short refresher on Probability

Training and Testing Error

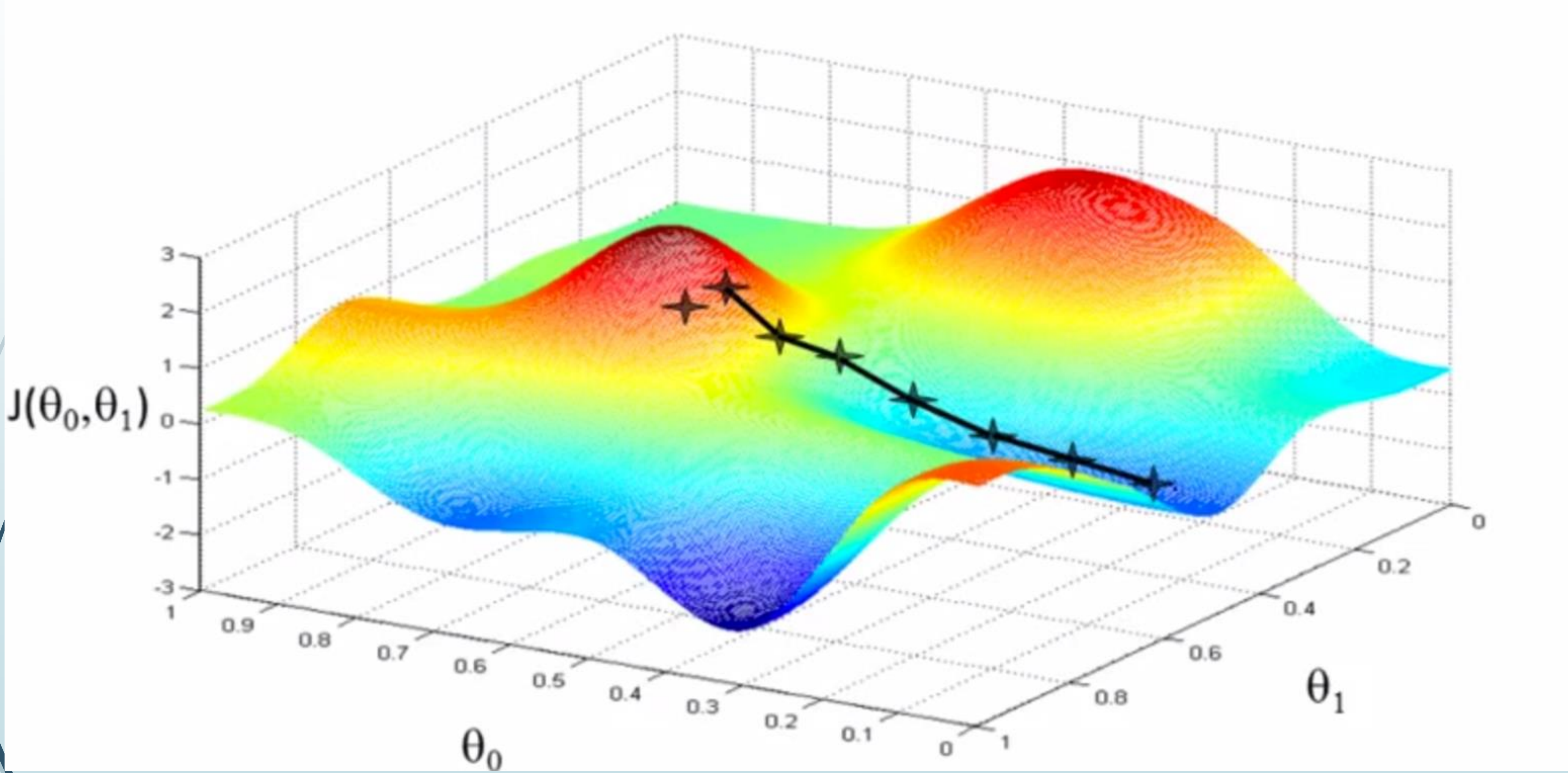


Overfitting symptom: Testing error >> Training error

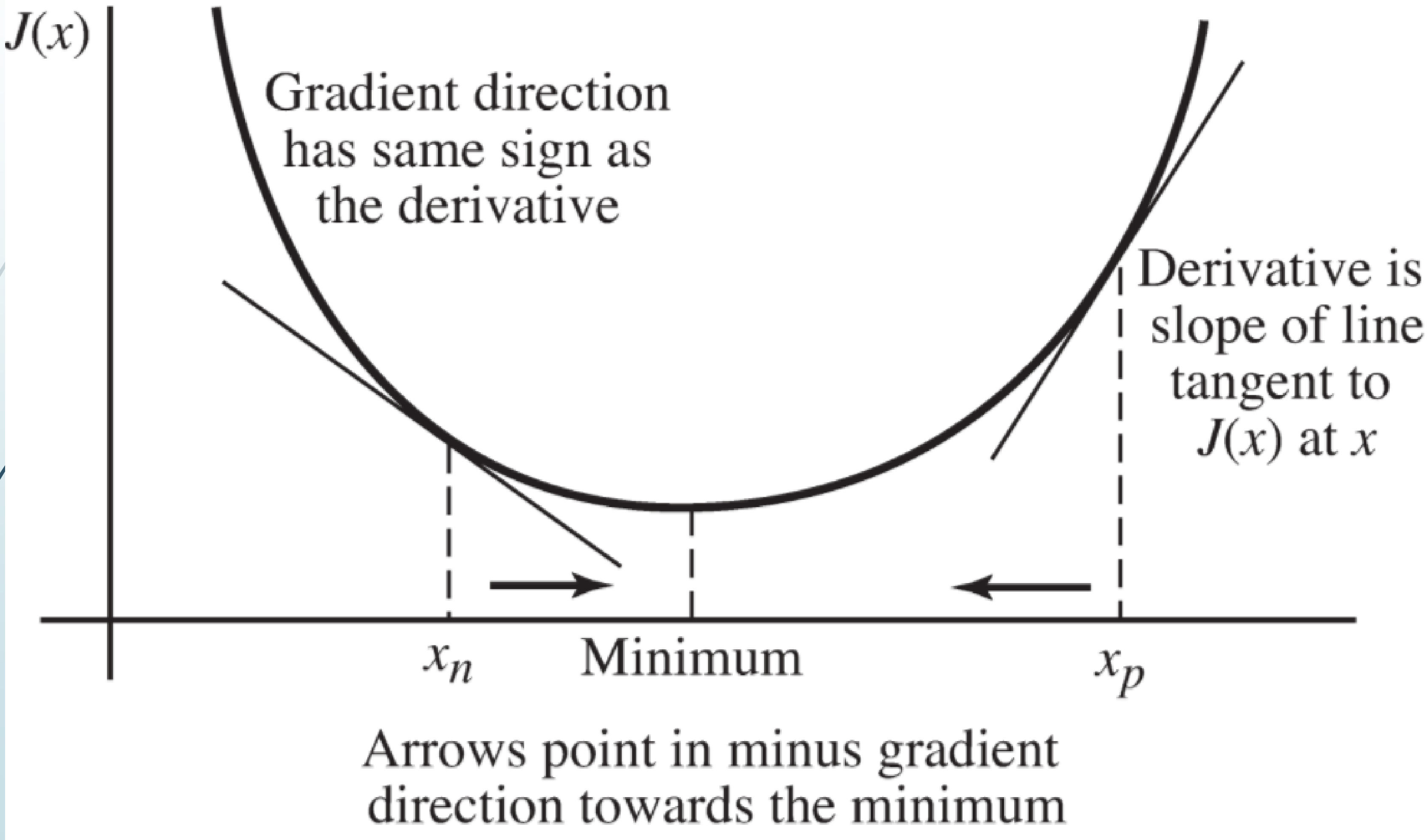
Linear regression

- We use a concept called Least square to find the model
  - **$y = a + b.x + c.x^2 + d.x^3$**
  - Above model (we have lots of x and y) and we want to learn the values of (a,b,c,d,..) the parameters of the model so that we can minimize the error between actual y and predicated y
  - Take a point(x0, y0), for a given (a,b,c,d…) we find predicted $y0_p$
  - Find the square of error $(y0 - y0_p)^2$
  - Add up the squared error terms for all sample points in training
  - Use some kind of algorithm to find the best possible value of (a,b,c,d…) which minimizes the "average squared error"

Gradient Descent



Gradient Descent - concept



Gradient Descent - ALgorithm

Gradient descent algorithm

repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \qquad (\text{for } j = 0 \text{ and } j = 1)$$

}

Correct: Simultaneous update

$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$
$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$
$\theta_0 := \text{temp0}$
$\theta_1 := \text{temp1}$

Linear Regression

- Quick intro to scikit
- Ingredients Bivariate data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- Model: $y_i = f(x_i) + E_i$

  where f (x) is some function, E_i random error.
- X_n Total squared error:

$$\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

- Model allows us to predict the value of y for any given value of x.
  - x is called the independent or predictor variable.
  - y is the dependent or response variable.

Examples of f(x)

- Lines: $y = ax+b+E$
- Polynomials: $y = ax^2+bx+c+E$
- Others: $y = a/x + b.x + E$
- Others: $y = a \sin(x) + b.e^x+c+E$

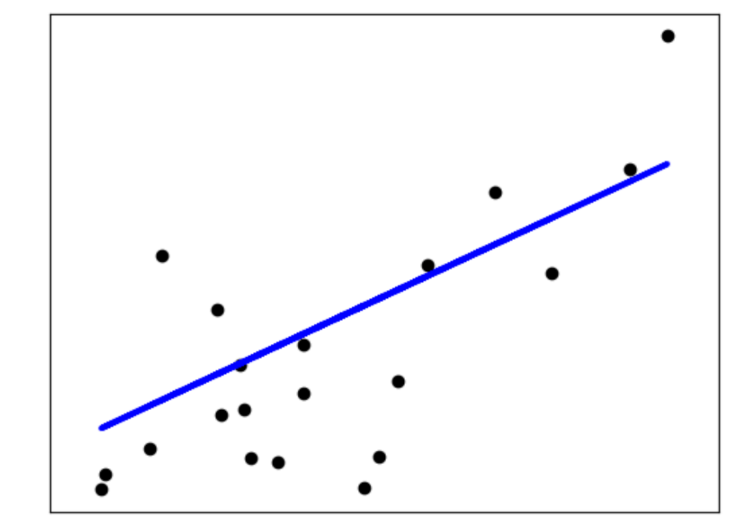
- Are these linear? Linear in what?

Session 3

- Case Studies – 15 mins
- Recap of Session 2
- Basic theory of Machine Learning
- Walkthrough of some supervised learning models

Supervised Learning - Linear/Logistic Regression

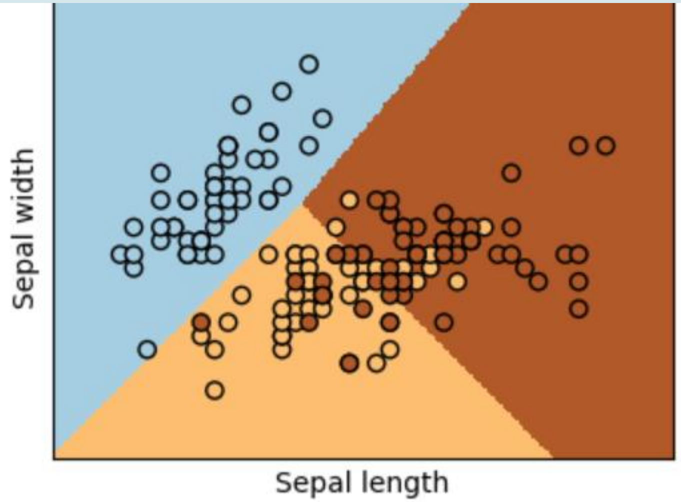
- X – training data
- Y – outcome [0,1], [0,1,2,3] etc – binary/categorical variable for classification and Y is continuous for Linear regression
- Model
  - $z = w^T . x$ w = [w_1, w_2, w_3, w_4, …] are the weights we need to learn
  - $\hat{y} = \frac{1}{1+e^{-z}}$ for classification
  - $\hat{y} = z$ for regression



- Metric/error we want to minimize using SGD
  - Classification: Cross entropy loss

    $-y.\ln(\hat{y}) - (1-y) * \ln(1-\hat{y})$:
  - Regression: Mean Squared loss

    $(y-\hat{y})^2$

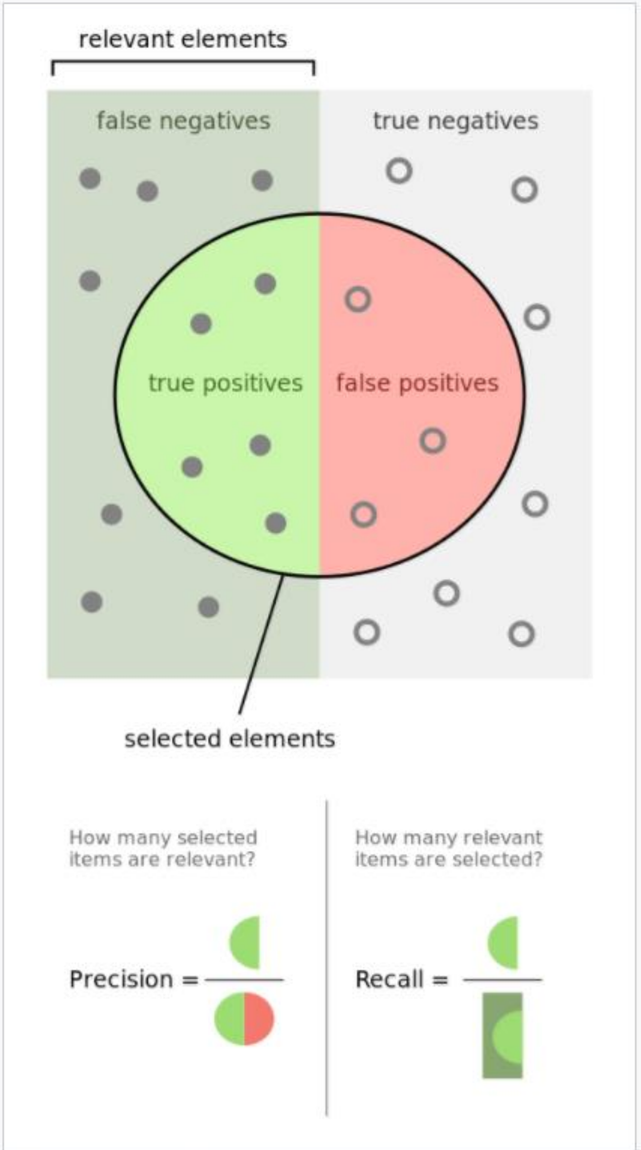


Classification - F1 score

Confusion Matrix

		Actual class	
		Cat	**Non-cat**
Predicted class	**Cat**	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	17 True Negatives

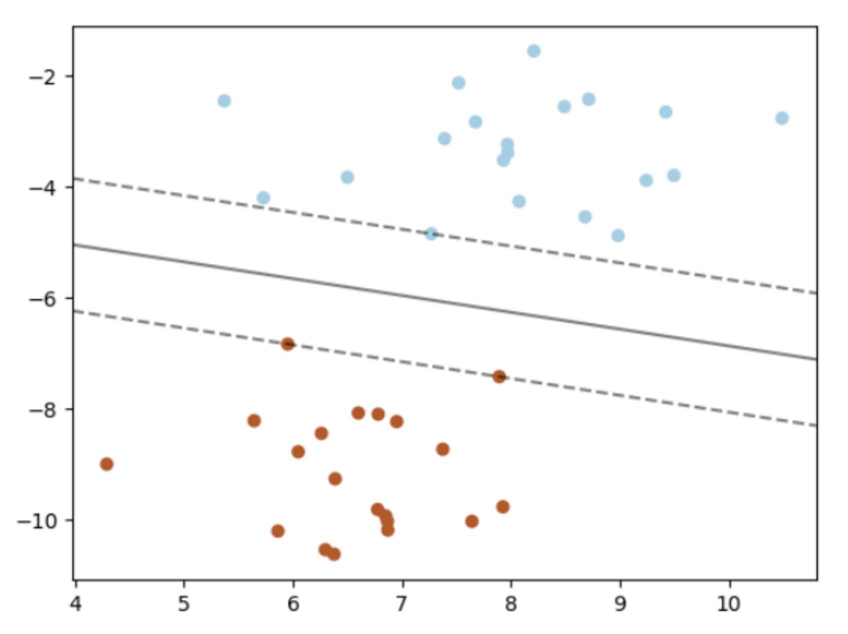
$$F_1 = 2 \cdot \frac{1}{\frac{1}{\mathrm{recall}} + \frac{1}{\mathrm{precision}}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$



Precision and recall

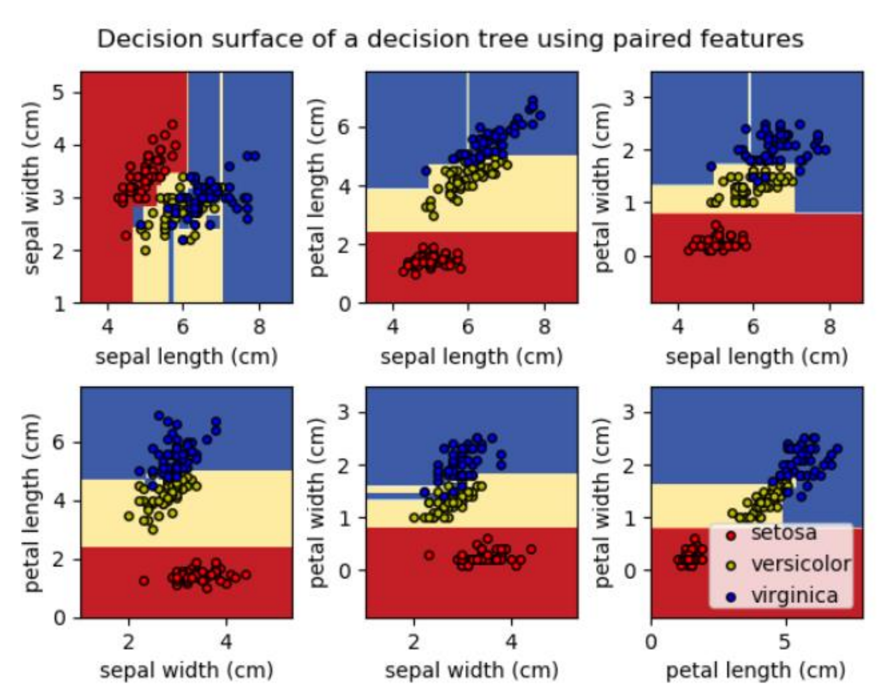
Supervised Learning - Classification – Support vector Classifiers

- Similar to Logistic classifier. The separating boundary is fine tuned to provide maximum separation between the line and nearest data points
- They fall into quadratic optimization techniques
- Have been most popular for complex problems till recently.



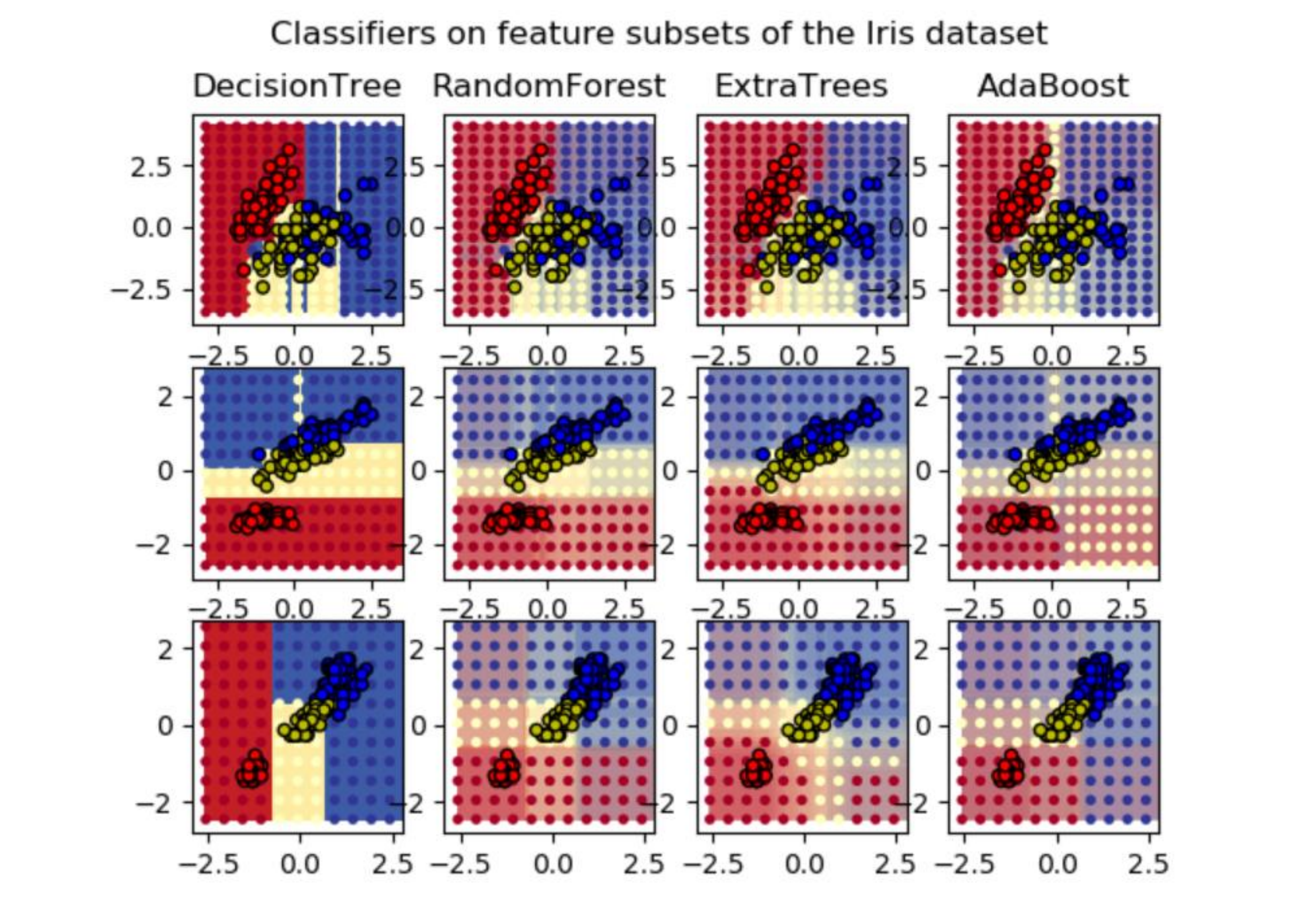
Supervised Learning - Classification – Decision Tree

- Decision Trees (DTs) are a non-parametric supervised learning method. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.
- X, y with y as categorical
- Model – non parametric
- Splits happen using
  - Ginni coefficient / Information gain
  - Algorithm - CART



Supervised Learning - Classification – Random Forests

- One Example of ensemble methods
  - Multiple decision tree classifiers are combined
  - each tree built from a sample drawn with replacement



Appendix A - Probaility

Where is probability?

- We use probability and statistical theory to keep all these errors close and be able to predict general characteristics based on sample data.
- All the models of estimation (regression, Classifier trees, SVMs, Neural Networks) use this concept in some way or other
- In other words we are trying to maximize the chances of successful prediction on unseen data using a model built on seen data

Probability in regression model

We are trying to find a relation between price(dependent) and house size(independent) from some sample data which we are hoping can the possible relationship given the kind of model we have chosen.

Linear regression

- We use a concept called Least square to find the model
  - **$y = a + b.x + c.x^2+d.x^3$**
  - Above model (we have lots of x and y) and we want to learn the values of (a,b,c,d,..) the parameters of the model so that we can minimize the error between actual y and predicated y
  - Take a point(x0, y0), for a given (a,b,c,d…) we find predicted $y0_p$
  - Find the square of error $(y0- y0_p)^2$
  - Add up the squared error terms for all sample points in training
  - Use some kind of algorithm to find the best possible value of (a,b,c,d…) which minimizes the "average squared error"

Where is probability??

- Question: what do we think minimizing squared error will give us a good model of prediction for unseen data
- When we "somehow" find right (a,b,c,d…) to minimize training error, we are
  - maximizing the chances that final values of (a,b,c,d…) will decrease the expected error on unseen data

Basic concepts/terms of Probability

- Sample space
  - (example of dice) {1,2,3,4,5,6}
- Events (some examples)
  - event that ‘1’ shows up on a throw
- Random variables
  - X – x is the face value of the dice i.e. x can take values from 1 to 6
  - P(X=3) = 1/6; P(x=3 or X=1) = 2/6
- Joint Probability
  - Let there be two dice X and Y
  - P(X = 1 and Y = 3) = ??

Basic concepts/terms of Probability(2)

- **Conditional Probability**
  - Let there be two dice
    - X the value of first dice {1,2,3,4,5,6}
    - Y the value of 2nd dice {1,2,3,4,5,6}
    - Z = X+Y i.e. {2,3,4,…12}
    - Can we find the probability that we saw Z=6, then what is the probability that X = 3
    - P(X=3 | Z=6) = ???
- **Independence**
  - P(X=3 | Y=3) == P(X=3)
  - When happening an event does not impact the chances of happening on other event
- **Basic rules of Conditional Probability**
  - P(X=x, Y=y) = P(X=x | Y=y) * P(Y=y)

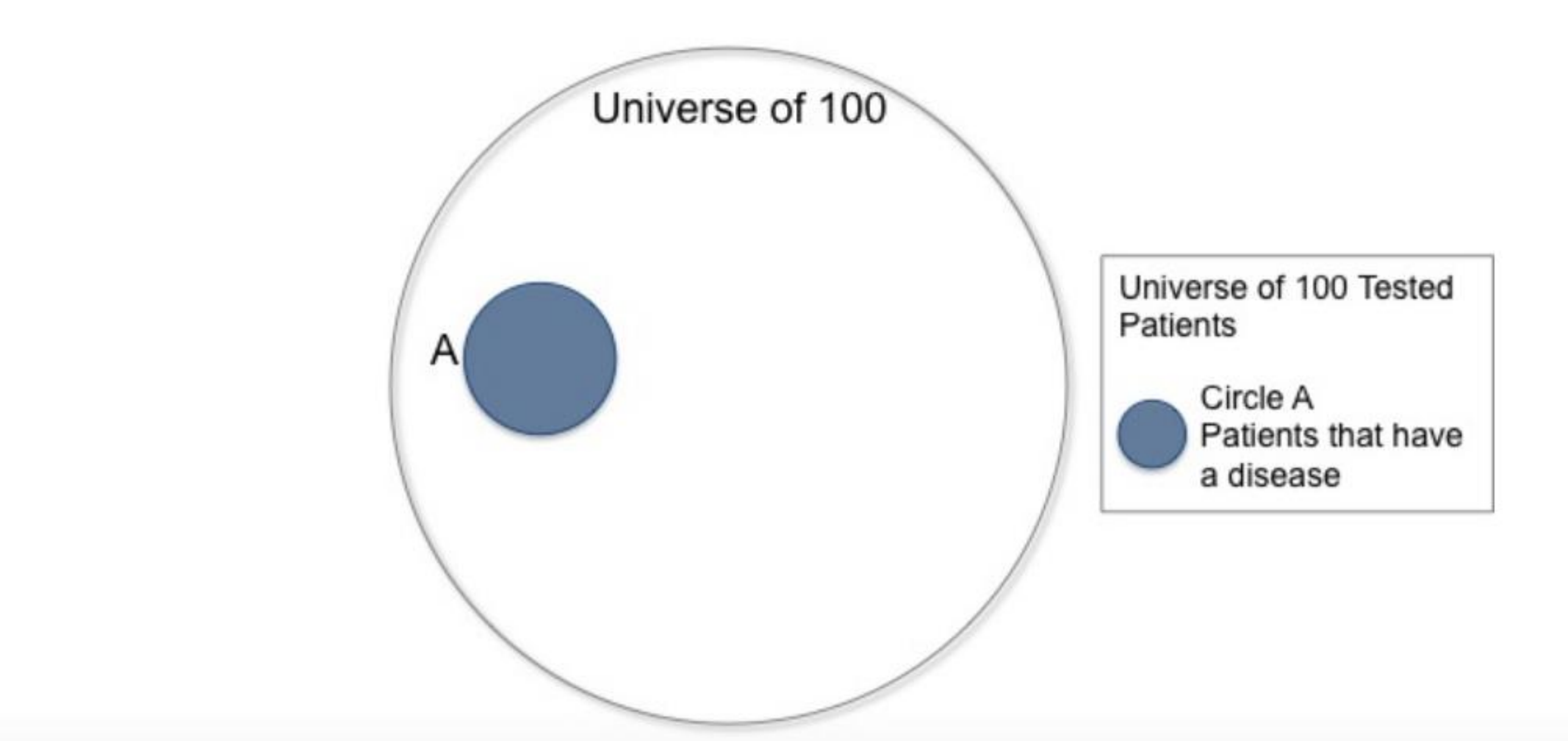
Basic concepts/terms of Probability(3)

- **Bayes Theorem Probability**
  - P(Y/X) = (P(X/Y)*(PY))/P(X)
  - P(X,Y) = P(X | Y) . P(Y) = P(Y | X) . P(X)

- A disease is present in 5 out of 100 people, and a test that is 90% accurate (meaning that the test produces the correct result in 90% of cases) is administered to 100 people. If one person in the group tests positive, what is the probability that this one person has the disease?

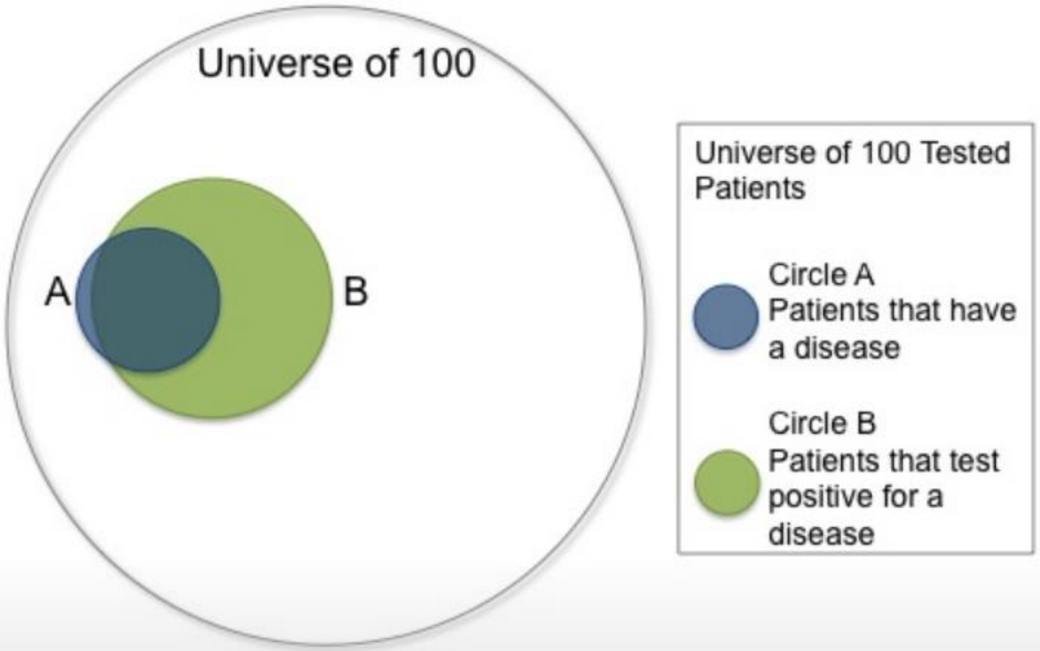
Basic concepts/terms of Probability(4)

- Circle A has 5 patients



Basic concepts/terms of Probability(5)

Next, overlay a circle to represent the people who get a positive result on the test. We know that 90% of those with the disease will get a positive result, so need to cover 90% of circle A, but we also know that 10% of the population who does not have the disease will get a positive result, so we need to cover 10% of the non-disease carrying population (the total universe of 100 less circle A).



Basic concepts/terms of Probability(6)

- Overlap of A&B = 4.5
- Total of B = 14
- So probability that a person tested +ve is actually ill = 4.5/14

- X = event that person is ill $P(X=1) = 0.95$
- Y = event that test is +ve $P(Y=1) = 0.9$
- We want to find $P(X=1 | Y=1)$
- We can use bayes theorem
  - $P(X=1 | Y=1) = P(Y=1 | X=1).P(X=1) \;/\; P(Y=1)$
  - $P(Y=1) = P(Y=1 | X=1).P(X=1) + P(Y=1 | X=0).\, P(X=0)$ (Total probability theorem)
  - $P(Y=1) = 0.9*0.05 + 0.1*0.95 = 0.045 + 0.095 = 0.14$
  - $P(X=1 | Y=1) = 0.045/0.14 = 4.5/14$ – same as above

Basic concepts/terms of Probability(7)

- Mean and variance
  - E[X] – i.e. what is the mean expected value of X
  - Example of single throw of a die
  - $Var(X) = E[X^2] - (E[X])^2$ – higher the variance more uncertain the outcome. Variance of zero means no uncertainty
- Some common distributions
  - Binomial
  - Multinomial
  - Poisson
  - Normal distribution