18CSE751 – Introduction to Machine Learning Lecture 4: Basic Statistics

Dr.Vani Vasudevan

Professor –CSE, NMIT

UNIT I

Introduction : Machine Learning, Types of Machine Learning, Machine Learning Process, Supervised Learning, Examples of Machine Learning Applications,

Machine Learning Preliminaries: Weight Space, Curse of Dimensionality, Testing Machine Learning Algorithms: Overfitting, Training, Testing, and Validation Sets, Confusion Matrix, Accuracy Metrics, ROC Curve, Unbalanced Datasets, Measurement Precision,

Basic Statistics : Averages, Variance, Covariance, Gaussian, Bias, Variance Tradeoff .

TYPES OF MACHINE LEARNING

- https://machinelearningmastery.com/types-of-learning-in-machine-learning/

---- Worthy read (given proper book references too)

EXERCISE: CONFUSION MATRIX

- Use the 4 x 4 confusion matrix and answer the following questions

1. What percent of the instances were correctly classified?

2. According to confusion matrix, how many *class 1/class 2 /class 3 /class 4* instances are there in the dataset?

1. How many instances were incorrectly classified with *class 1/ class2/class 3/class 4?*

2. Calculate sensitivity, specificity of class *1/class2/class 3/ class 4.*

1. Calculate FPR and FNR of *class 1/class2/class 3/class 4.*

2. Calculate F1 score class *1/class2/class 3/class 4.*

		Computed Decision			
		Class 1	Class 2	Class 3	Class 4
Actual	Class 1	10	3	3	2
	Class 2	3	20	0	1
	Class 3	2	2	15	2
	Class 4	3	3	1	20

OUTLINE

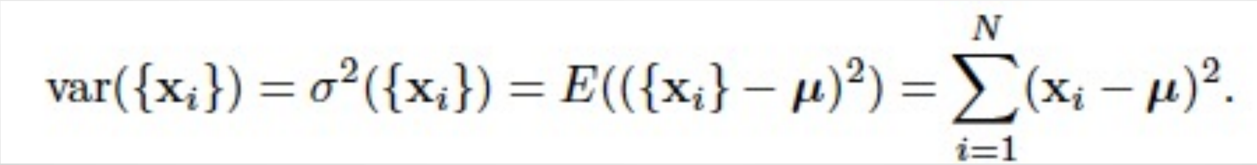
- BASIC STATISTICS
  - Average
  - Variance
  - Covariance
  - Gaussian
  - Bias , Variance trade-off

AVERAGE

- **Two numbers that can be used to characterise a dataset: the Mean and the Variance.**
- The **mean is the most used average** of a set of data and is the value that is found by adding up all the points in the dataset and dividing by the number of points.
- There are two **other averages that are used: the Median and the Mode.**
- The **median is the middle value,** so the most common way to find it is to sort the dataset according to size and then find the point that is in the middle.
- The **mode is the most common value,** so it just requires counting how many times each element appears and picking the most frequent one.

VARIANCE

- **The Variance of the set of numbers is a measure of how spread out the values are.**
- It is computed as the sum of the squared distances between each element in the set and the expected value of the set (the mean, μ):



$$\mathrm{var}(\{\mathbf{x}_i\}) = \sigma^2(\{\mathbf{x}_i\}) = E((\{\mathbf{x}_i\} - \mu)^2) = \sum_{i=1}^{N}(\mathbf{x}_i - \mu)^2.$$

- The square root of the variance, , is known as the **Standard Deviation.**

COVARIANCE…

- The **variance** looks at the **variation in one variable compared to its mean.**
- We can generalise this to look at **how two variables vary together,** which is known as the covariance.
- It is a measure of **how dependent the two variables are (in the statistical sense).** It is computed by:

$$\operatorname{cov}(\{\mathbf{x}_i\}, \{\mathbf{y}_i\}) = E(\{\mathbf{x}_i\} - \boldsymbol{\mu})E(\{\mathbf{y}_i\} - \boldsymbol{\nu}),$$

where v is the mean of set $\{y_i\}$.

COVARIANCE…

- If two variables are independent, then the **covariance is 0** (the variables are then known as uncorrelated),
- If they both increase and decrease at the same time, then the **covariance is positive,**
- and if one goes up while the other goes down, then the **covariance is negative.**
- The covariance can be used to look at the **correlation between all pairs** of variables within a set of data.

COVARIANCE…

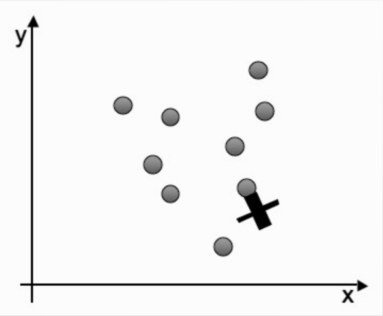
- Covariance of each pair are then put together into what is imaginatively known as the **covariance matrix**. It can be written

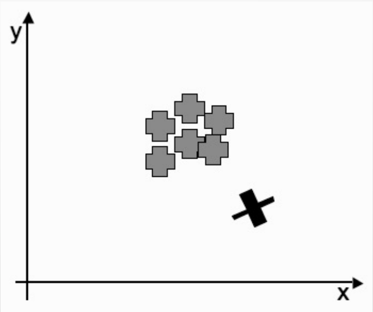
$$\boldsymbol{\Sigma} = \begin{pmatrix} E[(\mathbf{x}_1 - \boldsymbol{\mu}_1)(\mathbf{x}_1 - \boldsymbol{\mu}_1)] & E[(\mathbf{x}_1 - \boldsymbol{\mu}_1)(\mathbf{x}_2 - \boldsymbol{\mu}_2)] & \dots & E[(\mathbf{x}_1 - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_n)] \\ E[(\mathbf{x}_2 - \boldsymbol{\mu}_2)(\mathbf{x}_1 - \boldsymbol{\mu}_1)] & E[(\mathbf{x}_2 - \boldsymbol{\mu}_2)(\mathbf{x}_2 - \boldsymbol{\mu}_2)] & \dots & E[(\mathbf{x}_2 - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_n)] \\ \dots & \dots & \dots & \dots \\ E[(\mathbf{x}_n - \boldsymbol{\mu}_n)(\mathbf{x}_1 - \boldsymbol{\mu}_1)] & E[(\mathbf{x}_n - \boldsymbol{\mu}_n)(\mathbf{x}_2 - \boldsymbol{\mu}_2)] & \dots & E[(\mathbf{x}_n - \boldsymbol{\mu}_n)(\mathbf{x}_n - \boldsymbol{\mu}_n)] \end{pmatrix}$$

- Where x_i is a column vector describing the elements of the ith variable, and μ_i is their mean.
- Note that the matrix is square, that the elements on the leading diagonal of the matrix are equal to the variances, and that it is symmetric since $cov(xi, x_j) = cov(x_j\ , xi)$.
- It can also be written in matrix form as $\;= E[(X - E[X])(X - E[X])^T\ ]$, recalling that the mean of a variable X is E(X).

COVARIANCE…

- **Covariance Matrix says how the data varies along each data dimension.**
- This is useful if we want to think about distances again.
- Suppose consider two datasets shown in figure below, and the test point (labelled by the large ‘x’ in the figures) and asked you if the ‘x’ was part of the data.





COVARIANCE…

- Construct a distance measure that takes into account spread of the data . it is called the **Mahalanobis Distance**

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})},$$

where **x** is the data arranged as a column vector, **μ** is column vector representing the mean, and $\sum^{-1}$ is the inverse of the covariance matrix.

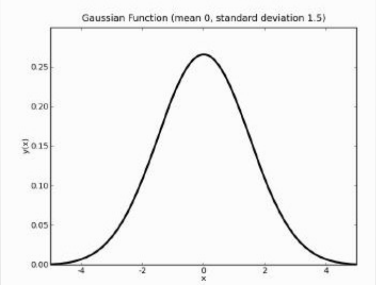
- If covariance matrix = Identity matrix, then the **Mahalanobis Distance reduces to the Euclidean Distance.**

Note : In NumPy there is a function that estimates the covariance matrix of a dataset np.Cov(x) for data matrix x and the inverse is called np.Linalg.Inv(x).

THE GAUSSIAN…

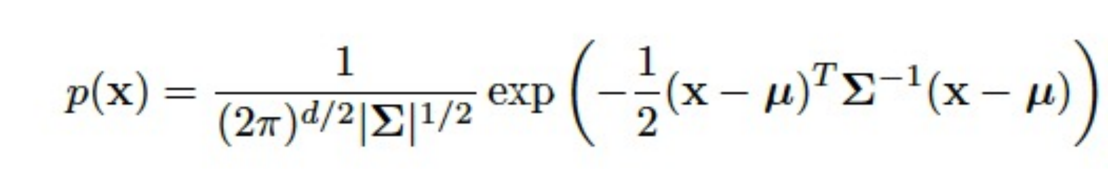
- THE GAUSSIAN(NORMAL): THE PROBABILITY DISTRIBUTION THAT IS MOST WELL KNOWN
- . IN ONE DIMENSION IT HAS THE FAMILIAR ‘BELL-SHAPED’ CURVE, AND ITS EQUATION IN ONE DIMENSION IS:

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



THE GAUSSIAN…

- The gaussian distribution turns up in many problems because of the central limit theorem, which says that lots of small random numbers will add up to something gaussian. In higher dimensions it looks like:

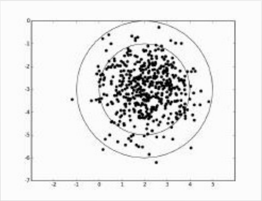


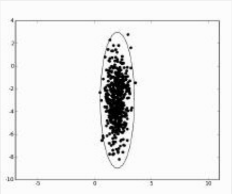
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

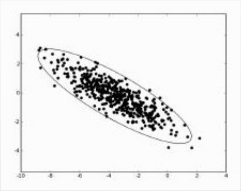
where ∑ is the $n\ x\ n$ covariance matrix (with I I being its determinant and −1 being its inverse).

THE GAUSSIAN

- The first case is known as a spherical covariance matrix, and has only 1 parameter.
- The second and third cases define ellipses in two dimensions, either aligned with the axes (with n parameters) or more generally, with n^2 parameters.







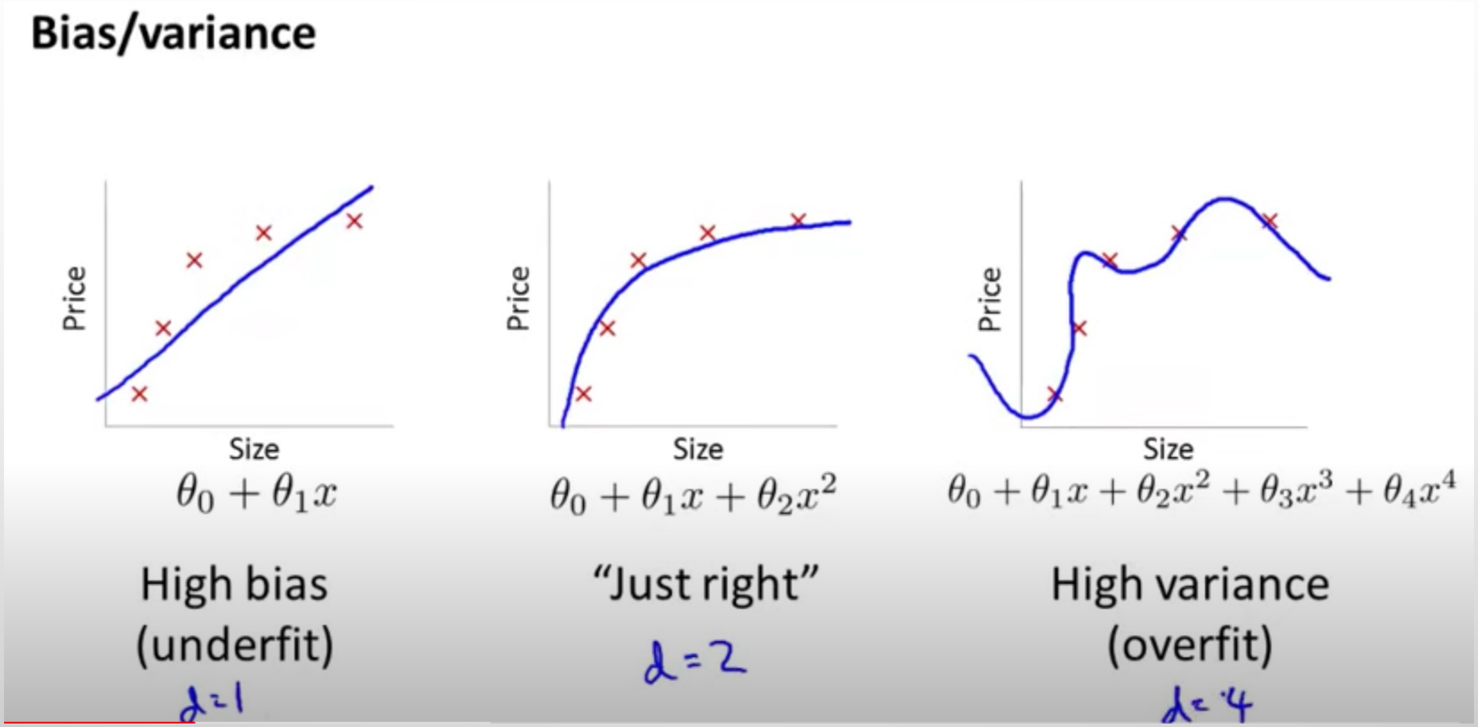
THE BIAS-VARIANCE TRADEOFF…

- More complicated models have inherent dangers such as **overfitting,** and **requiring more training data,** and we have seen the **need for validation data** to ensure that the model does not overfit.
- There is another way to understand this idea that **more complex models do not necessarily result in better results**. It is called **bias-variance trade-off.**

THE BIAS-VARIANCE TRADEOFF

- **A model can be bad for two different reasons.** Either it is **not accurate** and doesn't match the data well, or it is **not very precise** and there is a lot of variation in the results.
- The first of these is known as the bias, while the second is the statistical variance.
- More complex classifiers will tend to improve the bias, but the cost of this is higher variance, while making the model more specific by reducing the variance will increase the bias.

THE BIAS-VARIANCE TRADEOFF



Bias/variance

$\theta_0 + \theta_1 x$

High bias
(underfit)

$d=1$

$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

$d=2$

$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance
(overfit)

$d=4$

Source : Lecture 10.4 — Advice For Applying Machine Learning | Diagnosing Bias Vs Variance — [Andrew Ng]

UNIT I - SUMMARY

Introduction : Machine Learning, Types of Machine Learning, Machine Learning Process, Supervised Learning, Examples of Machine Learning Applications,

Machine Learning Preliminaries: Weight Space, Curse of Dimensionality, Testing Machine Learning Algorithms: Overfitting, Training, Testing, and Validation Sets, Confusion Matrix, Accuracy Metrics, ROC Curve, Unbalanced Datasets, Measurement Precision,

Basic Statistics : Averages, Variance, Covariance, Gaussian, Bias, Variance Tradeoff .

REFERENCES

1. STEPHAN MARSLAND**, MACHINE LEARNING, AN ALGORITHMIC PERSPECTIVE,** CRC PRESS SECOND EDITION, 2015.