MACHINE LEARNING

Shoaib Farooq

Department of Computer Science

Instructor Information

Instructor	Shoaib Farooq	E-mail	m.shoaib1050@gmail.com
GitHub	https://github.com/SHOAIB1050		
YouTube	https://www.youtube.com/c/pakithub		
LinkedIn	https://www.linkedin.com/in/shoaib-farooq-b5b190105/		

Let's Start .....

Lecture #5

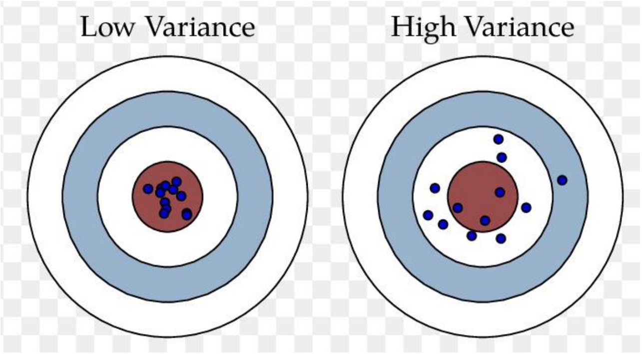
GOALS

- **This Lecture Will Cover:**
- **Gaussian Naive Bayes (GNB)**

BASICS STATISTICS

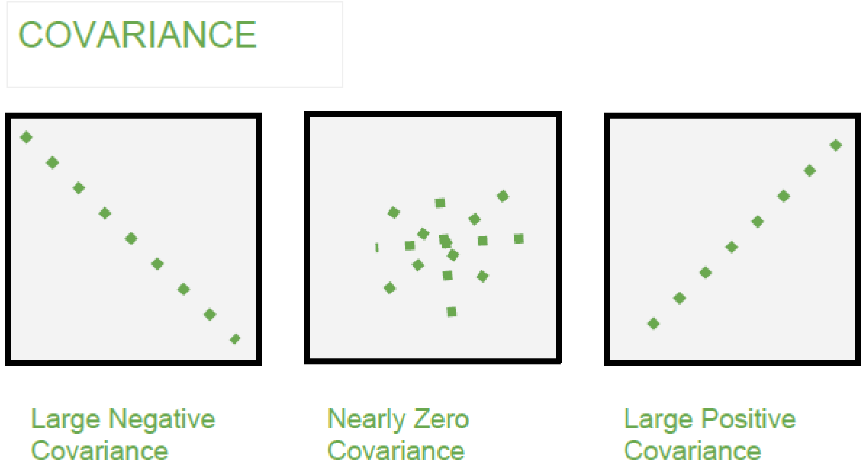
- Variance
  - It refers to the spread of a data.
  - Identifies how far each number in the data set is from the mean

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$$



BASICS STATISTICS

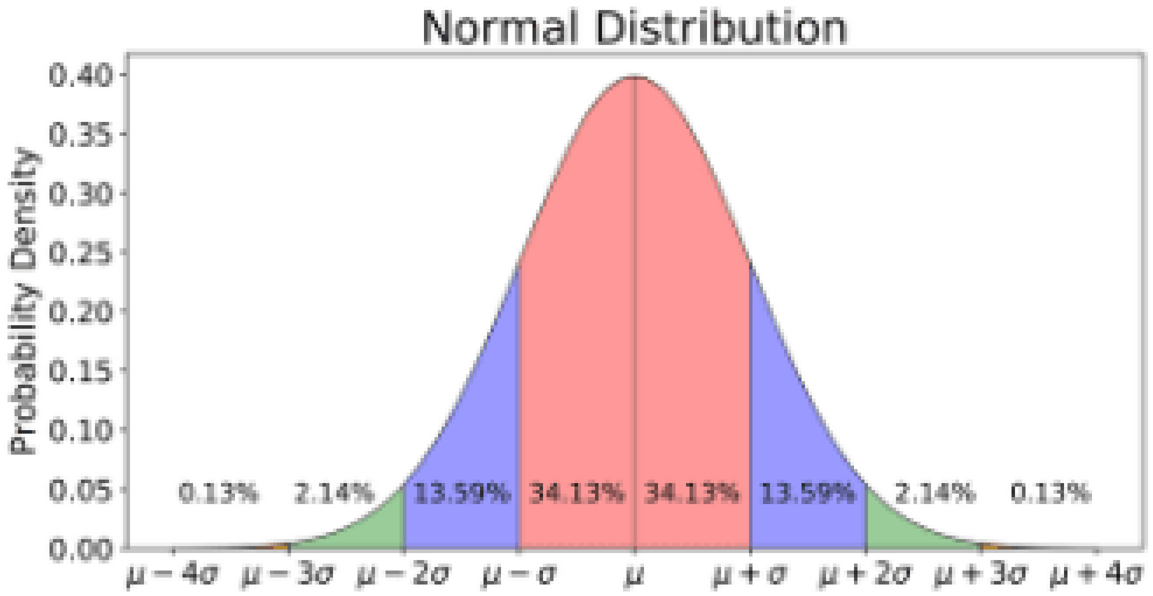
- Covariance
  - Covariance provides information about how the two random variables are related to one another.
  - Or it refers to the measure of how two random variables in a data set will change together.



$$COV(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

GAUSSIAN NAIVE BAYES (GNB)

Gaussian Naive Bayes (GNB) is a variant of the Naive Bayes algorithm used for classification problems. It assumes that the continuous features in the dataset are normally distributed (Gaussian distribution).



GAUSSIAN NAIVE BAYES (GNB)

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

Where:

- $P(y|X)$ is the posterior probability of class y given the feature vector X.
- $P(X|y)$ is the likelihood of the features given the class.
- $P(y)$ is the prior probability of class y.
- $P(X)$ is the evidence (a normalizing constant).

GAUSSIAN NAIVE BAYES (GNB)

GNB assumes that each feature follows a Gaussian (normal) distribution for each class. The probability density function (PDF) for a Gaussian distribution is:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Where:

- x_i is the feature value.
- μ_y and σ_y are the mean and standard deviation of the feature for class y.

STEPS TO APPLY GAUSSIAN NAIVE BAYES

1. **Training:**
   - Calculate the mean μ_y and variance σ_y^2 for each feature, for every class, based on the training data.
2. **Prediction:**
   - For each new instance, compute the likelihood for every class using the Gaussian PDF.
   - Multiply the likelihoods by the prior probabilities and select the class with the highest posterior probability.

ADVANTAGES OF GAUSSIAN NAIVE BAYES

- **Fast and Efficient:** It requires relatively less computational power, especially for large datasets.
- **Simple to Implement:** The assumptions and computations are straightforward.
- **Works Well For Continuous Data:** Since it assumes the data is normally distributed, it's well-suited for continuous variables.

LIMITATIONS OF GAUSSIAN NAIVE BAYES

- **Strong independence assumption:** In real-world scenarios, features are often correlated, which violates the assumption and may affect performance.
- **Sensitive to the distribution assumption:** If the data does not follow a Gaussian distribution, the performance may degrade.

DATASET

Person	Height (ft)	Weight (lbs)	Foot size (inches)
Male	6.00	180	12
Male	5.92	190	11
Male	5.58	170	12
Male	5.92	165	10
Female	5.00	100	6
Female	5.50	150	8
Female	5.42	130	7
Female	5.75	150	9

DATASET

Person	Height (ft)	Weight (lbs)	Foot size (inches)
Male	6.00	180	12
Male	5.92	190	11
Male	5.58	170	12
Male	5.92	165	10
Female	5.00	100	6
Female	5.50	150	8
Female	5.42	130	7
Female	5.75	150	9

$$P(Male) = 4/8 = 0.5$$

$$P(Female) = 4/8 = 0.5$$

DATASET

Person	Height (ft)	Weight (lbs)	Foot size (inches)
Male	6.00	180	12
Male	5.92	190	11
Male	5.58	170	12
Male	5.92	165	10
Female	5.00	100	6
Female	5.50	150	8
Female	5.42	130	7
Female	5.75	150	9

$$P(Male) = 4/8 = 0.5$$

$$P(Female) = 4/8 = 0.5$$

Male:

$$Mean\ (Height) = \frac{(6+5.92+5.58+5.92)}{4} = 5.855$$

$$Variance\ (Height) = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$= \frac{(6-5.855)^2+(5.92-5.855)^2+(5.58-5.855)^2+(5.92-5.855)^2}{4-1}$$

$$= 0.035055$$

DATASET

Person	Height (ft)	Weight (lbs)	Foot size (inches)
Male	6.00	180	12
Male	5.92	190	11
Male	5.58	170	12
Male	5.92	165	10
Female	5.00	100	6
Female	5.50	150	8
Female	5.42	130	7
Female	5.75	150	9

$$P(Male) = 4/8 = 0.5$$

$$P(Female) = 4/8 = 0.5$$

Male:

$$Mean\ (Height) = \frac{(6+5.92+5.58+5.92)}{4} = 5.855$$

$$Variance\ (Height) = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$= \frac{(6-5.855)^2+(5.92-5.855)^2+(5.58-5.855)^2+(5.92-5.855)^2}{4-1}$$

$$= 0.035055$$

Sex	Mean (height)	Variance (height)	Mean (weight)	Variance (weight)	Mean(foot size)	Variance (foot size)
Male	5.855	0.035033	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667

DATASET

Sex	Mean (height)	Variance (height)	Mean (weight)	Variance (weight)	Mean(foot size)	Variance (foot size)
Male	5.855	0.035033	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667

New Instance to be Classified is:

Sex	Height(ft)	Weight(lbs)	Foot size(inch)
Sample	6	130	8

$$P(Male) = 4/8 = 0.5$$
$$P(Female) = 4/8 = 0.5$$

DATASET

Sex	Mean (height)	Variance (height)	Mean (weight)	Variance (weight)	Mean(foot size)	Variance (foot size)
Male	5.855	0.035033	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667

New Instance to be Classified is:

Sex	Height(ft)	Weight(lbs)	Foot size(inch)
Sample	6	130	8

$$P(Male) = 4/8 = 0.5$$

$$P(Female) = 4/8 = 0.5$$

$$Posterior\ (Male) = \frac{P(M) * P(H|M) * P(W|M) * P(FS|M)}{Evidence}$$

$$Posterior\ (Female) = \frac{P(F) * P(H|F) * P(W|F) * P(FS|F)}{Evidence}$$

DATASET

Sex	Mean (height)	Variance (height)	Mean (weight)	Variance (weight)	Mean(foot size)	Variance (foot size)
Male	5.855	0.035033	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667

New Instance to be Classified is:

Sex	Height(ft)	Weight(lbs)	Foot size(inch)
Sample	6	130	8

$$P(Male) = 4/8 = 0.5$$

$$P(Female) = 4/8 = 0.5$$

$$Posterior\ (Male) = \frac{P(M) * P(H|M) * P(W|M) * P(FS|M)}{Evidence}$$

$$Posterior\ (Female) = \frac{P(F) * P(H|F) * P(W|F) * P(FS|F)}{Evidence}$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$P(H|M) = \frac{1}{\sqrt{2 * 3.142 * 0.035033}} * e^{-\frac{(6-5.855)^2}{2*0.035033}} = 1.5789$$

$$P(W|M) = 5.9881e^{-6}$$

$$P(FS|M) = 1.3112e^{-3}$$

$$P(H|F) = 2.2346e^{-1}$$

$$P(W|F) = 1.6789e^{-2}$$

$$P(FS|F) = 2.8669e^{-1}$$

DATASET

Sex	Mean (height)	Variance (height)	Mean (weight)	Variance (weight)	Mean(foot size)	Variance (foot size)
Male	5.855	0.035033	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667

New Instance to be Classified is:

Sex	Height(ft)	Weight(lbs)	Foot size(inch)
Sample	6	130	8

$$P(Male) = 4/8 = 0.5$$

$$P(Female) = 4/8 = 0.5$$

$$P(H|M) = \frac{1}{\sqrt{2 * 3.142 * 0.035033}} * e^{-\frac{(6-5.855)^2}{2*0.035033}} = 1.5789$$

$$P(W|M) = 5.9881e^{-6}$$

$$P(FS|M) = 1.3112e^{-3}$$

$$P(H|F) = 2.2346e^{-1}$$

$$P(W|F) = 1.6789e^{-2}$$

$$P(FS|F) = 2.8669e^{-1}$$

$$Posterior\ (Male) = \frac{P(M)*P(H|M)*P(W|M)*P(FS|M)}{Evidence} = 0.5 * 1.5789 * 5.9881e^{-6} * 1.3112e^{-3} = 6.1984e^{-9}$$

$$Posterior\ (Female) = \frac{P(F)*P(H|F)*P(W|F)*P(FS|F)}{Evidence} = 0.5 * 2.2346e^{-1} * 1.6789e^{-2} * 2.8669e^{-1} = 5.377e^{-4}$$

DATASET

Sex	Mean (height)	Variance (height)	Mean (weight)	Variance (weight)	Mean(foot size)	Variance (foot size)
Male	5.855	0.035033	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667

New Instance to be Classified is:

Sex	Height(ft)	Weight(lbs)	Foot size(inch)
Sample	6	130	8

$$P(Male) = 4/8 = 0.5$$
$$P(Female) = 4/8 = 0.5$$

$$P(H|M) = \frac{1}{\sqrt{2 * 3.142 * 0.035033}} * e^{-\frac{(6-5.855)^2}{2*0.035033}} = 1.5789$$
$$P(W|M) = 5.9881e^{-6}$$
$$P(FS|M) = 1.3112e^{-3}$$

$$P(H|F) = 2.2346e^{-1}$$
$$P(W|F) = 1.6789e^{-2}$$
$$P(FS|F) = 2.8669e^{-1}$$

Female

$$Posterior\ (Male) = \frac{P(M)*P(H|M)*P(W|M)*P(FS|M)}{Evidence} = 0.5 * 1.5789 * 5.9881e^{-6} * 1.3112e^{-3} = 6.1984e^{-9}$$

$$Posterior\ (Female) = \frac{P(F)*P(H|F)*P(W|F)*P(FS|F)}{Evidence} = 0.5 * 2.2346e^{-1} * 1.6789e^{-2} * 2.8669e^{-1} = 5.377e^{-4}$$

Thank You ☺