AIT 736 – Home Work 1

Vishnu Lasya Marthala – G01275122

**Introduction to Applied PATTERN RECOGNITION AND MACHINE LEARNING**

1. **Describe supervised and unsupervised learning and provide one specific example for both types of learning. For each example state the problem and describe the evaluation plan (training, validation and test sets).**

According to Arthur Samuel, Machine Learning is a field that gives the computer or the machine the ability to learn without being programmed explicitly. According to Tom Mitchell, A computer program is said to learn from the experience E w.r.t some Task T and performance measure P, if performance improves with experience. We have different kind of learning algorithms and the two main types are Supervised Learning Algorithms and Unsupervised Learning Algorithms.

**Supervised Learning**, the name itself says that this algorithm is guided or supervised with the help of the labled training data and target data (i.e., the right result/answer for the input(s) is given). The aim of the supervised learning is to forecast the outcomes. There are namely two types of supervised learning. They are Regression and Classification. Linear Regression and Logistic Regression come under Regression Algorithms whereas the Decision Trees, K Nearest Neighbours and Random Forest come under Classification problems. Some of the practical applications of Supervised Learning are Evaluation of Risk, Forecasting sales or profits in the Business sector, predicting the house prices, predicting how many items will get sold over a period of time, Predicting if Cancer is benign or malignant, predicting if the computer is hacked or not etc., In order to go ahead with supervised learning, one needs to collect the training data (having the target vectors), there is feature representation and the goal is to find develop a model that finds the best hypothesis. The main components of Supervised learning are Task (predicting the target variable), Experience (which is nothing but the training set) and the Performance metric (that is one must know how to evaluate your algorithm by comparing with other algorithms and this is where real world is more interested in). Evaluating the model is more like making sure the unseen, new data is correctly classified i.e., there should be generalization. We can also find the accuracy on the test data for evaluating the model.

**Example:** Classification Problem – KNN (K Nearest Neighbours) -Distance based Algorithm

Classifying the person having the tumour which can be Benign or Malignant.

**Task**: To find if the person has benign tumour or malignant tumour

**Experience:** Having the trained data that correctly has the classified tumour type

**Performance Metrics:** The commonly used technique to measure the performance of the classification problems is Confusion matrix which consists of the True Positives, False Positives, True Negatives and False Negatives. We can compute Accuracy, Precision, Recall and F-Score from the confusion matrix data. Accuracy is the most commonly used performance metric which is nothing but the ratio of number of correctly classified observations to the total number of observations evaluated. We can also use ROC curve, Cross validation for evaluation (**Evaluation Plan**).

**Unsupervised Learning**, the name itself says that this algorithm is **not** guided or supervised with the help of target data (i.e., the right result/answer for the input(s) is given). There is no labelled data in unsupervised learning. The aim of the unsupervised learning is to discover the underlying patterns hidden behind the data. There are namely two types of unsupervised learning. They are Clustering and Association. K-means Clustering come under Clustering Algorithms whereas the Apriori and Association Rule Mining come under Association problems. Some of the practical applications of Unsupervised Learning are Recommendation Systems and Anomaly detection like detecting the Credit Card Fraud etc., In supervised learning we are told the answer (target vector) explicitly but in unsupervised learning we are just given the data with no labels ( or all the observations under same label). In this learning we are given the dataset but we are not told what to do with it exactly, we are also not told what each data point represents. We are just told to find some structure or pattern in the given data. Density Estimation and Dimensionality Reduction also comes under the Unsupervised Learning. Google News is able to display the new stories related to the same topic together at one place with the help of the Unsupervised Learning. Understanding Genomics, organizing the different computer machine clusters in large data centers, Analyzing the social network (facebook, Instagram), Astronomical Data Analysis which gives useful theories of how galaxies are formed etc., are all possible only because of unsupervised learning. There is also a cocktail party algorithm under unsupervised machine learning which is able to separate the overlapping audio recordings that are mixed/added together.When compared to the supervised learning, unsupervised learning has fewer tests and models and creates a less controllable environment as the computer itself is creating the outcomes for us.

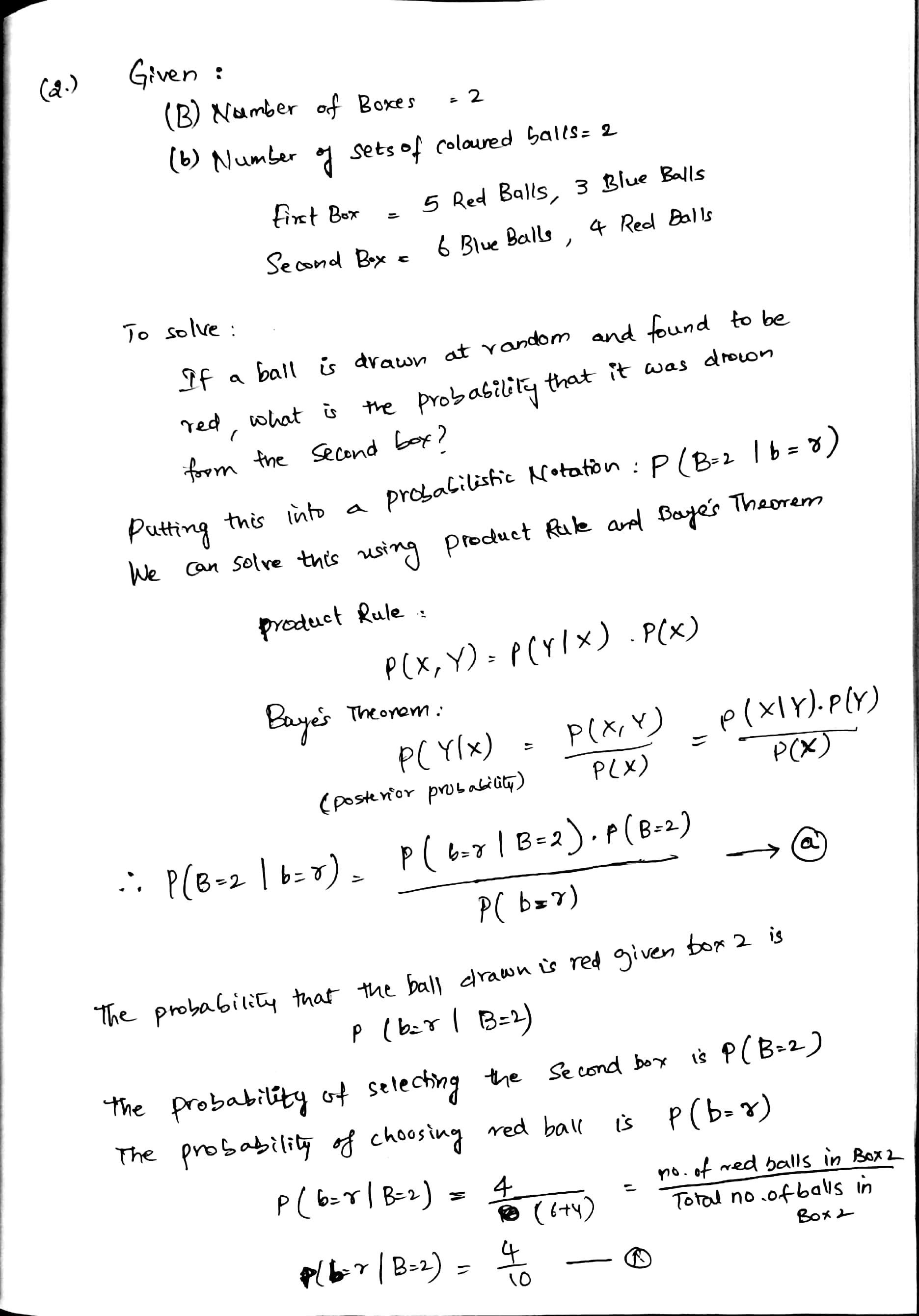
**Example:** Clustering Problem – K-means Clustering Algorithm

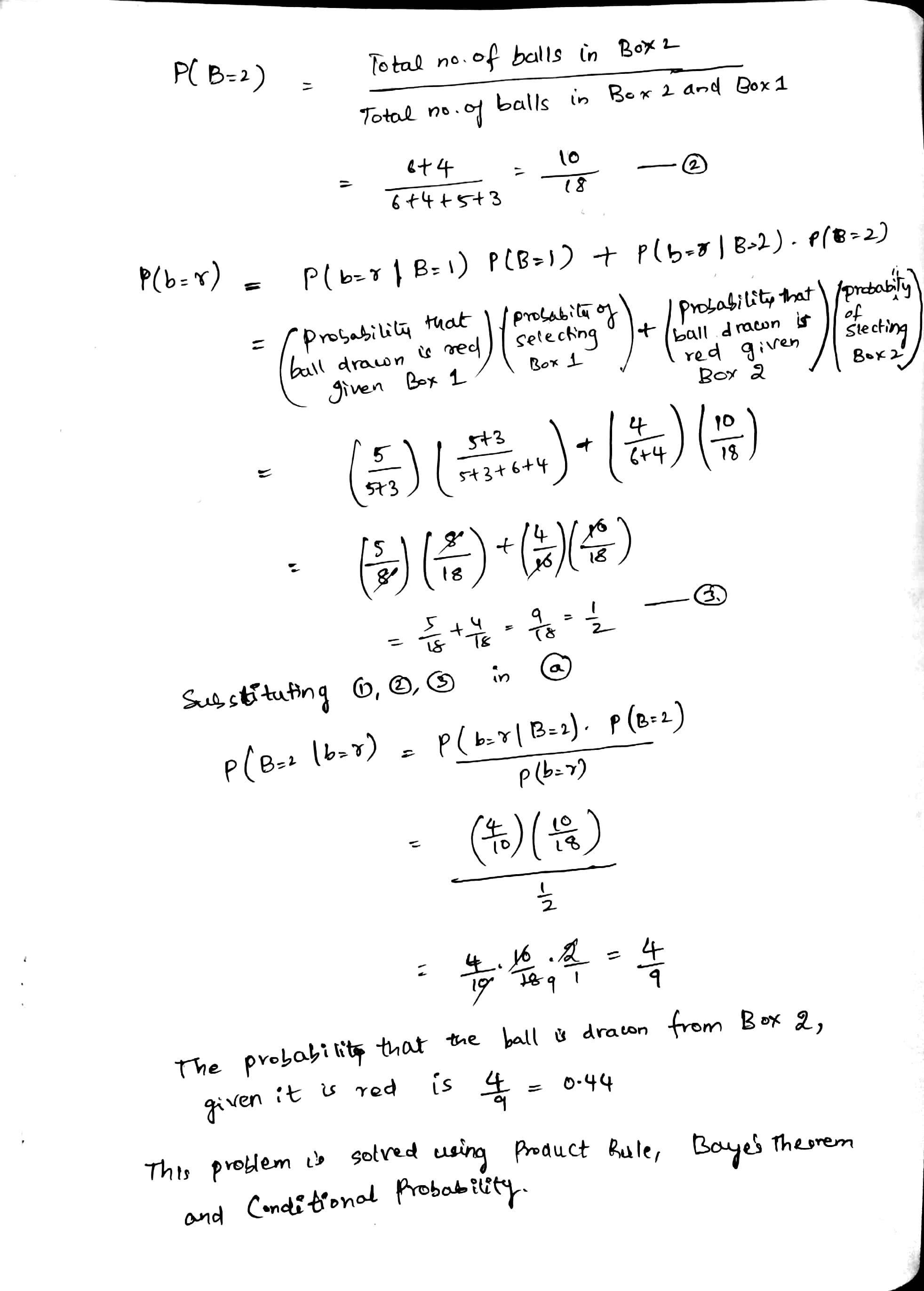
How is Google News able to display the new stories related to the same topic together at one place (one cluster for one topic) with the help of the Unsupervised Learning

**Task**: To find the clusters of different news topics accurately

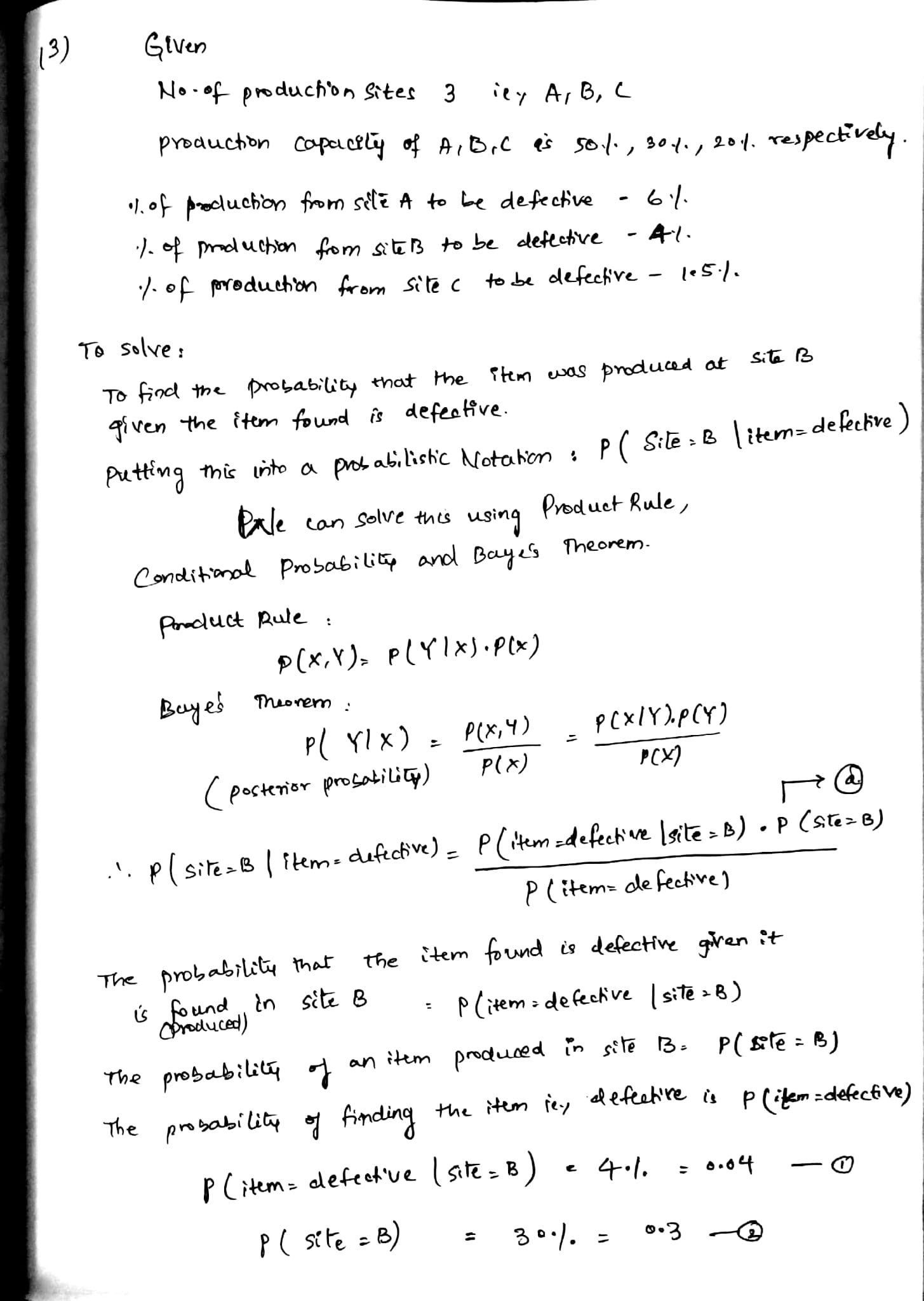
**Performance Metrics:** The commonly used technique to measure the performance of the clustering problems where we validate the correct number of clusters needed for the task to be solved. To evaluate the appropriate correct number of clusters, we have two methods. One is Elbow method and the other is Silhouette Coefficient. For the elbow method we need Sum of Squared Error (SSE). The place where the SSE curve starts to bend is known as elbow point. Silhouette Coefficient can be used to check if the cluster count from the elbow method can be considered or not. This values ranges between 1 and -1. This is calculated based on the inter and intra distances between the clusters. (**Evaluation Plan**).

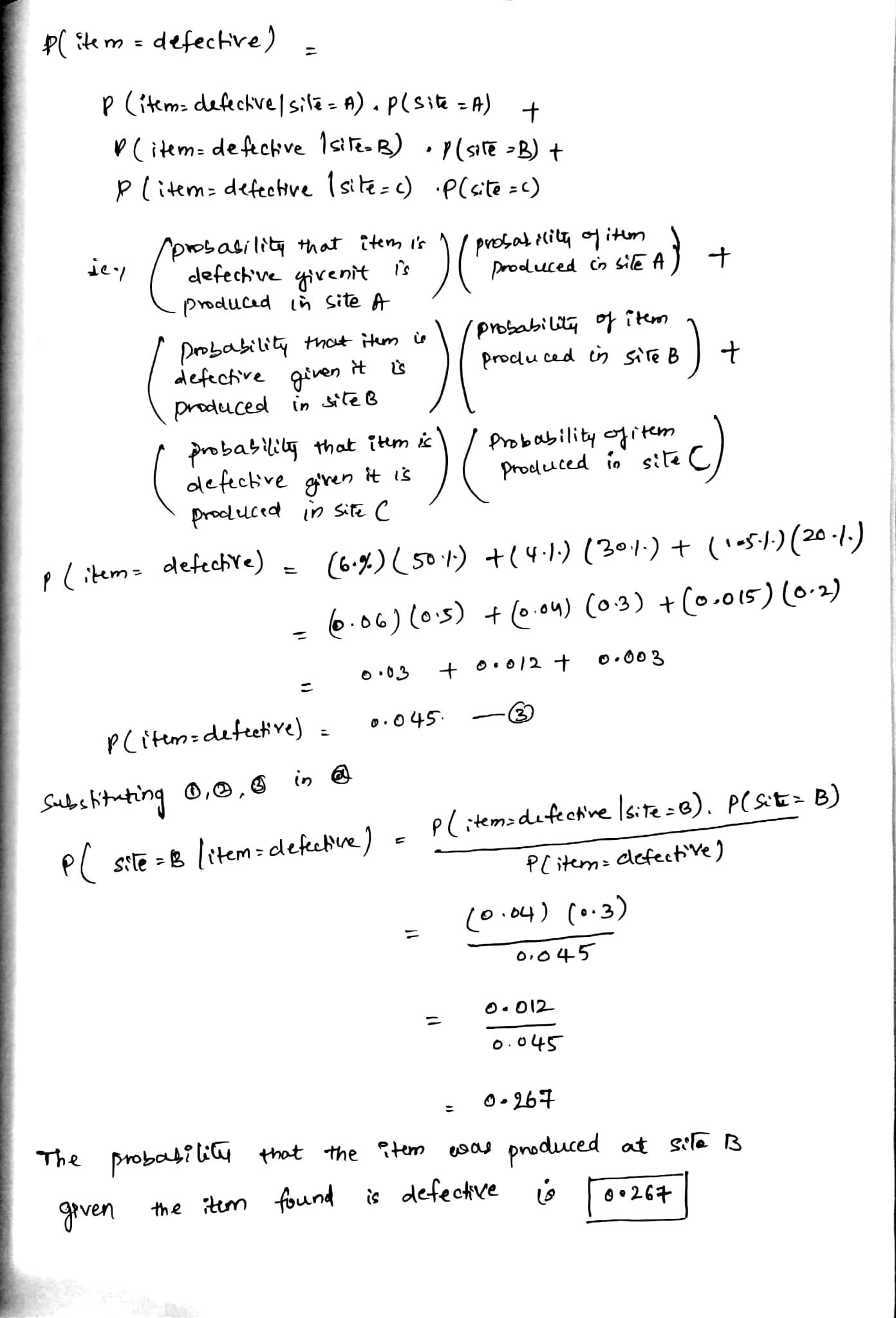
1. **There are 2 boxes containing 2 sets of colored balls. The first box contains 5 red balls and 3 blue balls, while the second box contains 6 blue balls and 4 red balls. If a ball is drawn at random and found to be red, what is the probability that it was drawn from the second box?**

****

****

1. **A production company has 3 sites A, B, and C, with a production capacity of 50%, 30%, and 20% respectively. 6% of total production from site A is found to be defective, 4% of total production from site B is defective, and 1.5% of total production from site C is defective. An item selected at random is found to be defective. Find the probability that the item was produced at site B.**

****

****

**4. Explanation for KNN AND CNN:**

When k=1, there will be only one point in the convex area to classify the new unseen data. So, for example if the new data point is classified as letter ‘A’ when k=1. The same unseen data point can be classified as a different letter when k=5 is used because now we will be having 5 data points to choose from the convex area created by the algorithm i.e., to compare with and take the majority vote using Euclidian distance. When k=1, it is only compared with one data point but when k=5 it compares with 5 data points. So as the k-value increases and when k=n is used then the entire training data will be used to for every prediction and the unseen data is just predicted as the majority class in the dataset. So, we should select the k-value which gives the minimum error rate which is estimated with the help of test set. K-nearest neighbours doesn’t work well especially when the features are on different scale. In our example all the features are on the same scale i.e., between 0 to 15. As the k value increase we get more reasonable decision boundary. Instance based learning just compares the unseen data with training data and it doesn’t actually come up with the model. It just stores the data. KNN can be very slow if we do not apply any techniques like sorting/ clustering/ condensing. Because classification time is the main drawback when KNN is used and especially when the dataset is very large. Also, when there are more features in the dataset and then Euclidian distance uses many attributes and there will be little distance between the different samples as there are many attributes. So relevant feature selection is very important when KNN is used. The advantage is KNN can model more complex hypothesis as there is no specific geometric shape of boundary. As said, KNN works fast if we apply specific techniques like Condensed KNN, which mean it reduces the samples of training set accordingly and then KNN is applied. As in k=1 in KNN we got the accuracy rate as 95% but when CNN with 1 neighbour is used we got can accuracy of 78%. There is a reduced accuracy because the training sample size is reduced here in CNN. But still, CNN can give a pretty good approximation (instead of explicit generalization) of the new unseen data within less computational search time and it also reduces the memory size issue.

**Sources Used:**

1. Class Presentations and Materials
2. <https://realpython.com/k-means-clustering-python/#understanding-the-k-means-algorithm>
3. <https://www.youtube.com/watch?v=PPLop4L2eGk&list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN&index=1>