1. Logistic Regression
   1. x will vote for party A if

x will vote for party B if

x will vote for party A or B randomly when

* 1. The threshold T of the probability p must first be defined to obtain a decision boundary of a more general form

1. Logistic Regression
   1. z will approach argmax where zj approaches 0 for all j except the argmax will approach 1
   2. In contrast, z will approach the same uniform value, i.e., 1/K
   3. Case 1 ():  
      Case 2 ():  
      Combining the two cases, we have
2. Feedforward Neural Network  
   We can regard this new variant of classification problem as a five binary classification problem instead of five-class classification problem.  
   1. While there may be many choices to modify the original network to handle this variant of the classification problem, one simple way could be to utilize the original activation function and , but modify only to use logistic regression for binary classification as mentioned previously, such as  
      This step is sufficient to solve the new variant of problem given that we have the objective function below to perform five binary classification.
   2. Following the modification in a, we could change the original loss function, which is sometimes called categorical cross-entropy loss, defined in lecture notes to the (Bernoulli or binary) cross entropy loss used in previous lecture notes in logistic regression, such as
   3. After defining a new network for five binary classification problem, we can simply use the to classify x into class i given that is greater than 0.5 (as the default threshold). Otherwise, we will indicate ‘none of the above’ or ‘reject’.
3. Convolutional Neural Networks
   1. Conv (128, 32, 7, 7, 2)

      2. Conv1 (128, 16, 1, 1, 1)
      3. Conv1 (128, 16, 1, 1, 1)
      4. Conv2 (16, 32, 7, 7, 2)
      5. Conv2 (16, 32, 7, 7, 2)
      6. Conv1 + Conv2
4. Principal Component Analysis
   1. 100%. In other words, the variance is completely preserved. By using two principal components obtained by PCA to project S linearly onto another two-dimensional space spanned by the principal components together, 100% of the total variance can be explained by the two principal components together since S also in a two-dimensional space. Thus, there is no dimensionality reduction and it is equivalent to linear transformation that maps S to the same space.
   2. In PCA, the total variance is the sum of the variance of the principal component which is the eigenvalue as shown in the lecture notes. The maximum percentage of total variance that can be explained is expressed by the sum of eigenvalues of the used principal components divided by the sum of eigenvalues of all principal components. By using one principal components, the maximum percentage of total variance is .
5. Clustering – Partitional Clustering
   1. For simplicity, we partition S into k clusters or k new sets of data points  
      which contains all data points that belongs to cluster i according to b with i = 1, …, k. By rewriting the objective function, we have  
      Minimizing the original objective function is equivalent to minimizing the mean squared distance for each individual cluster given by  
      Taking the derivative, we have  
      Thus, which corresponds to the cluster mean obtained after updating according to the local optimization algorithm.
   2. Denote S1 and S2 as the cluster 1 and cluster 2 with mean m1 and m2 respectively.  
      S1 = {(x, y) | 1<x<9} S2 = {(x, y) | 11<x<17} m1 = (5, 1.5) m2 = (14, 1.5)  
      S1 = {(x, y) | 1<x<9} S2 = {(x, y) | 11<x<17} m1 = (5, 1.5) m2 = (14, 1.5)  
      🡺 stop  
      S1 has 10 data points while S2 has 8 data points.
6. Clustering – Hierarchical Clustering