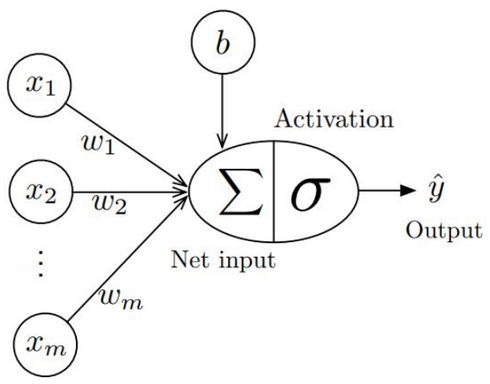
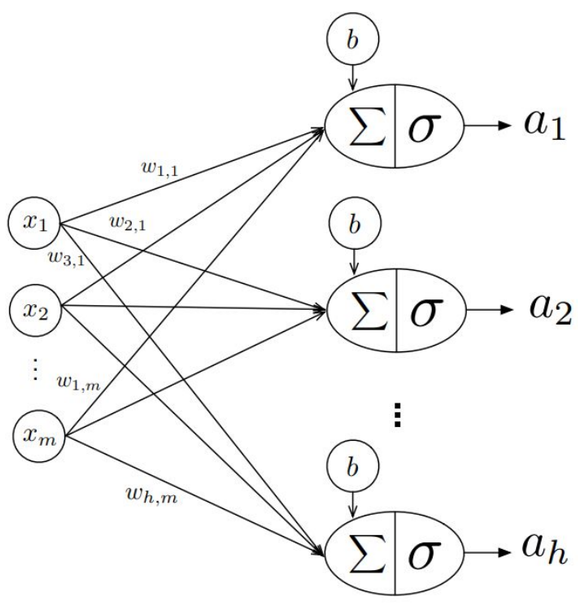
National Institute of Science Education and Research

CS 460

Anuleho Biswas (1911030)

SOFTMAX REGRESSION





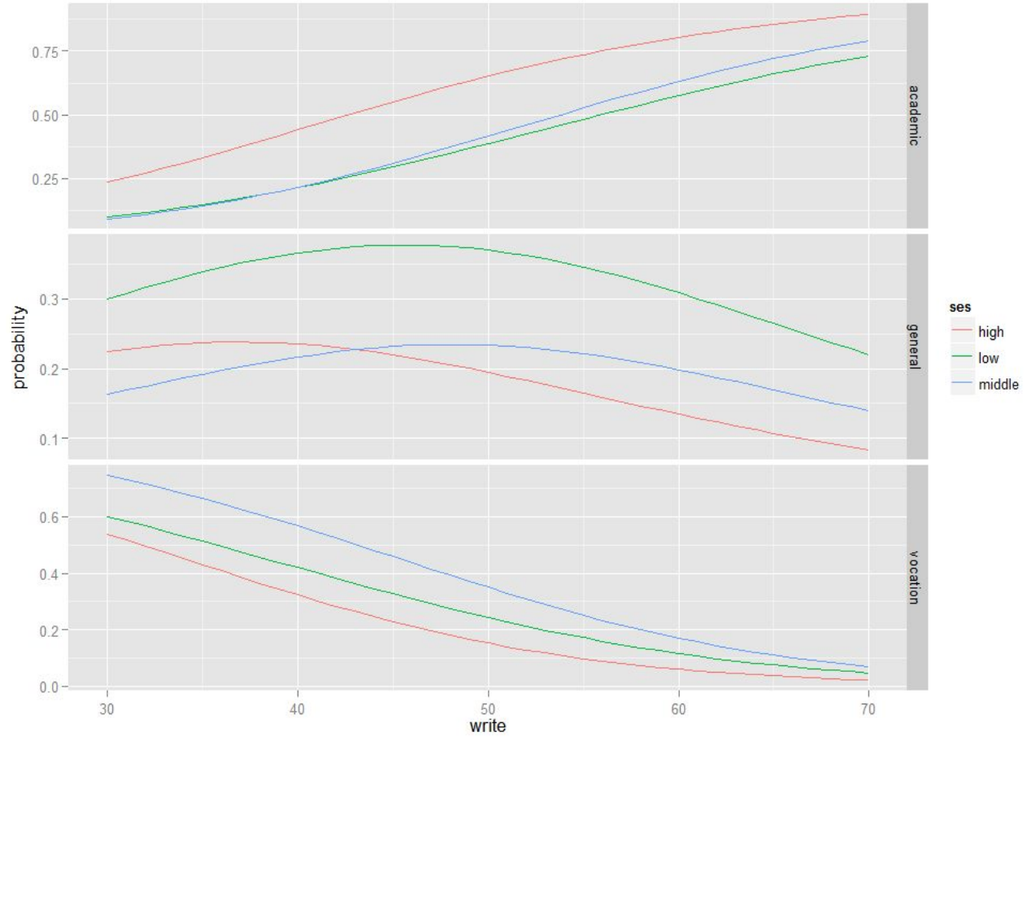
Binary logistic regression	***Multinomial logistic regression***
Sigmoid Function	**Softmax Function**
Maximum Likelihood Estimator	**Cross-Entropy Loss Function**
Gradient Descent	**Stochastic Gradient Descent**

Example:

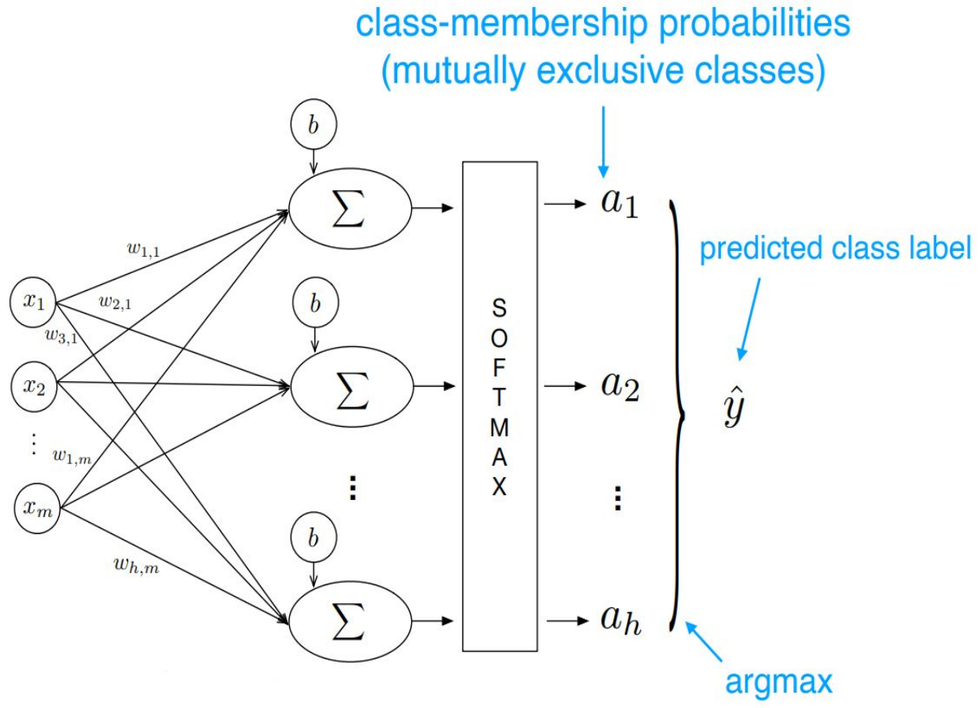
hsbdemo data set
The outcome variable is **prog**: program type.
The predictor variables are social **ses:** economic status,
A three-level categorical variable and **write:** writing score, a continuous variable.

Students have the option to choose from **general program**, **vocational program** and **academic program**. We cam model their choice, using their writing score (**write**) and their social economic status(**ses)**.

Result:



MLR has many steps common with **binary logistic regression**, and the only difference is the function for each step. In case of latter, Sigmoid function is used because it is a binary classification problem. In MLR, we use the **Softmax function** as it has more than two class.



Softmax follow the procedure to distribute probabilities for each output node. Our loss function rely on the activation function so loss function will be different because Activation function is different in Multinomial Logistic Regression.

$$P(y = t \mid z_t^{[i]}) = \sigma_{\text{softmax}}(z_t^{[i]}) = \frac{e^{z_t^{[i]}}}{\sum_{j=1}^{h} e^{z_j^{[i]}}}$$

$t \in \{j...h\}$

h is the number of class labels

Here regular **gradient descent** is being used instead **stochastic gradient descent** because of the excess number of features in our data. Calculating gradient descent for each feature would take significant amount of computation resource.

We are using **one-hot encoder** to transform the original values into one-hot encoded values because our predicted values are probabilities.

Application:

Natural language processing :

Multinomial Logistic Regression is an alternative to naive Bayes classifiers as it do not consider statistical independence of the random variables (in common ML term it is called *features*) . However, learning is slower than for a naive Bayes classifier, and thus it is not suitable for a very large number of classes for learning. Specifically, learning in a Naive Bayes classifier is a relatively simple matter (because it only include summation of number of co-occurrences of features and classes), But for a maximum entropy classifier the weights, which are typically maximized using maximum a posteriori (MAP) estimation, it use a iterative procedure for learning.

Medical Application :

There are many places applying this method in allergology. For instance, Mukesi et al. (2017) found out the most common reason of many types of skin allergies while Ranciere et al. (2013) studied the associations between dry night cough, atopy and allergic morbidity

Reference:

1] Mukesi, M., Phillipus, I. N., Moyo, S. R., & Mtambo O. P. L. (2017). Prevalence of Skin Allergies in Adolescents in Namibia. International Journal of Allergy Medications, 3(1):022.

2] Can, V. V. (2013). Estimation of travel mode choice for domestic tourists to Nha Trang using the multinomial probit model. Transportation Research Part A: Policy and Practice, 49, 149–159

3] https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/

4] Logistic regression - Binary, ordinal and multinomial

5]https://sebastianraschka.com/pdf/lecture-notes/stat453ss21/L08_logistic__slides.pdf