# Weighted Machine Learning Methods for Binary Classification Using Mutual Information

Tingfang Wang

**Abstract**

In machine learning, the conventional classification algorithms treat all the features equally without taking into consideration the fact that different attributes can have different levels of influence on a class. In this project, I propose an attribute weighting method based on mutual information and apply this method to three classical machine learning models for classification. The performance of the weighting method is evaluated by conducting experiments and comparing the classification accuracy on the Wisconsin Breast Cancer dataset. The weighted attribute model consistently outperformed or at least matched the performance of the corresponding conventional machine learning models in binary classification across all three models.

## 1 Introduction

In standard binary classification algorithms, all features are typically assigned equal importance. This overlooks the reality that certain attributes may exert varying degrees of influence on a class. Attribute weighting modifications are employed in machine learning models to enhance performance. In this project, I propose an attribute weighting method based on normalized mutual information, which is applied to three classical machine learning models for classification: **naive Bayes**, **KNN** and **logistic regression** [1].

The rest of this report is organized as follows. Section 2 describes the methodology of the three traditional classifiers and the corresponding weighting models in machine learning. Section 3 presents the implementation, including a description of the dataset and the results. Finally, the discussion is presented in section 4.

## 2 Methodology

### 2.1 Mutual Information

Mutual information (MI) measures the mutual independence between two random variables [2]. More specifically, it measures how much information we can learn about one random variable by observing the other random variable.

For two discrete random variables $X$ and $Y$, the mutual information between them can be calculated as:

$$\text{MI}(X, Y) = \sum_{x} \sum_{y} P(x, y) log(\frac{P(x, y)}{P(x)P(y)})$$

For two continuous random variables $X$ and $Y$, the mutual information between them can be calculated as:

$$\text{MI}(X,Y) = \int_y \int_x P(x,y) log(\frac{P(x,y)}{P(x)P(y)}) dxdy$$

where $P(x,y)$ is the joint probability mass/density function of $X$ and $Y$, and $P(x)$ and $P(y)$ are the marginal probability mass/density functions of $X$ and $Y$ respectively.

## 2.2 Normalized Mutual Information as Weight

The normalized mutual information (NMI) between $X$ and $Y$ is:

$$\text{NMI}(X,Y) = \frac{\text{MI}(X,Y)}{\text{mean}(H(X), H(Y))}$$

where $H(X)$ and $H(Y)$ are information entropy of $X$ and, $Y$ respectively.

Based on this, the weight of $i^{th}$ feature is defined as the NMI between the feature $X_i$ and the label $C$:

$$W(X_i) = \text{NMI}(X_i, C) = \frac{\text{MI}(X_i, C)}{\text{mean}(H(X_i), H(C))} \tag{1}$$

This definition is used for all the weighted classifiers in this project.

## 2.3 Weighted naive Bayes Classifier

In this project, I extend the conventional naive Bayes classifier to a weighted naive Bayes classifier by weighting the features when calculating the posterior probabilities. The weighted naive Bayes classifier is presented in the following equations.

The weighted posterior probability of the feature $X_i = x_i$ given the label $C_k$ is defined as:

$$P_W(x_i|C_k) = W(X_i)P(x_i|C_k)$$

where the weights are defined as in equation (1).

And, the weighted naive Bayes classifier for binary classification can be defined as:

$$\hat{c} = \arg\max_{k \in 1,2} P(C_k) \prod_{i=1}^{p} P_W(x_i|C_k) \tag{2}$$

While the original naive Bayes classifier for binary classification is defined as:

$$\hat{c} = \arg\max_{k \in 1,2} P(C_k) \prod_{i=1}^{p} P(x_i|C_k) \tag{3}$$

## 2.4 Weighted KNN

In the KNN classifier, the weighting method is integrated into kNN by defining a new weighted distance function. The weighted kNN computes the distance between an instance $X$ and a query $q$ using:

$$d_W(X_i, q) = (\sum_{i=1}^{p} W(X_i)\delta(X_i, q_i)^2)^{1/2} \tag{4}$$

Where $\delta(X_i, q_i) = |X_i - q_i|, i = 1, ..., p$ and the weights are defined as in equation (1).

Denote the new set of q's k-nearest neighbors found by the weighted distance function $d_W()$ as $N_{WK}(q)$. Given a query instance $q$, the predicted class label of this query is determined by:

$$\hat{c} = \arg\max_{C_k} \sum_{C_i \in N_{WK}(q)} I(C_i = C_k) \tag{5}$$

where $I(C_i = C_k)$ is an indicator function which yields 1 if the argument is true.

While the original KNN is defined by:

$$\hat{c} = \arg\max_{C_k} \sum_{C_i \in N_K(q)} I(C_i = C_k) \tag{6}$$

where $N_K$ is the set of $q's$ k-nearest neighbors found by the original Euclidean distance function $d()$.

## 2.5 Weighted logistic regression

For the weighted logistic regression, I incorporate the weights by raising each feature to the power of the respective weight obtained from mutual information. The weighted logistic regression is presented in the following equations.

The probabilities of the labels given the weighted feature set $x = x_1, ...x_p$ are defined as:

$$P_W(C = 1|x) = \frac{\exp(\beta_0 + \beta_1 {x_1}^{W(X_1)} + ... + \beta_p {x_p}^{W(X_p)})}{1 + \exp(\beta_0 + \beta_1 {x_1}^{W(X_1)} + ... + \beta_p {x_p}^{W(X_p)})}$$

and

$$P_W(C = -1|x) = \frac{1}{1 + \exp(\beta_0 + \beta_1 {x_1}^{W(X_1)} + ... + \beta_p {x_p}^{W(X_p)})}$$

where the weights are defined as in equation (1).

And, the weighted logistic regression for binary classification can be defined as:

$$\beta = \arg\max_{\beta} \prod_{i=1}^{N} P_W(C^{(i)}|x^{(i)}, \beta) \tag{7}$$

While the original logistic regression for binary classification is defined as:

$$\beta = \arg\max_{\beta} \prod_{i=1}^{N} P(C^{(i)}|x^{(i)}, \beta) \tag{8}$$

where $\beta = \{\beta_0, ..., \beta_p\}$, and

$$P(C = 1|x) = \frac{\exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}$$

$$P(C = -1|x) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}$$

# 3 Implementation

## 3.1 Description of the Dataset

The Wisconsin Breast Cancer (WBC) database was obtained from the University of Wisconsin Hospitals, Madison by Dr. William H. Wolberg [3]. This database contains 699 instances, of which 241 are malignant and 458 are benign. The original database has 16 instances which have some missing values. After removing these observations from the original database, 683 instances are remaining. The reduced database has 239 malignant instances and 444 benign instances.

The dataset contains the following features: (1) Clump thickness; (2) Uniformity of cell size; (3) Uniformity of cell shape; (4) Marginal adhesion; (5) Single epithelial cell size; (6) Bare nuclei; (7) Bland chromatin; (8) Normal nucleoli; and (9) Mitoses. The class labels are defined as 2 = benign, and 4 = malignant.

## 3.2 Performance Evaluation

The performance evaluation of the weighted naive Bayes and weighted logistic regression are carried out based on cross-validation, and the average accuracies across different folds are calculated for performance measurement. Several cross-validations with different folds, 3-fold to 11-fold, are applied in the experiments.

Table 1: Classification Accuracies (%) for Naive Bayes

| Cross Validation | 3cv | 5cv | 7cv | 9cv | 10cv | 12cv | Average |
|---|---|---|---|---|---|---|---|
| Naive Bayes | 97.00 | 97.15 | 97.15 | 97.00 | 97.16 | 97.02 | 97.08 |
| Weighted Naive Bayes | 97.00 | 97.15 | 97.15 | 97.00 | 0.9716 | 0.9702 | 97.08 |

In Table 1, the classification accuracies achieved by the naive Bayes classifier across different cross-validation runs are shown. Both the conventional and weighted naive Bayes models exhibit similar performances, averaging an accuracy of 97.08% across various cross-validations.

Table 2: Classification Accuracies (%) for Logistic Regression

| Cross Validation | 3cv | 5cv | 7cv | 9cv | 10cv | 12cv | Average |
|---|---|---|---|---|---|---|---|
| Logistic Regression | 95.35 | 95.36 | 95.81 | 96.26 | 95.81 | 95.81 | 95.73 |
| Weighted Logistic Regression | 95.35 | 96.41 | 95.96 | 96.40 | 96.41 | 96.41 | 96.16 |

The same comparison table is presented for the logistic regression. Table 2 demonstrates that the weighted logistic regression model either slightly outperforms or performs equally well compared to the conventional model in every simulation. This trend holds for the overall average classification accuracy as well.

For the weighted KNN model, instead of using cross-validation, the experiments are conducted based on different k values, 1, 3, 5, 7, and 9. The accuracies of the different k values, as well as the average accuracy for KNN and weighted KNN, are compared.

Table 3: Classification Accuracies (%) for KNN

| K | 1 | 3 | 5 | 7 | 9 | Average |
|---|---|---|---|---|---|---|
| KNN | 88.29 | 89.49 | 90.39 | 90.39 | 89.79 | 89.67 |
| Weighted KNN | 90.99 | 92.49 | 92.19 | 91.89 | 91.89 | 91.89 |

Table 3 illustrates the correct classification rates for different KNN models using varied k values. The highest accuracy achieved by the conventional KNN model was 90.39%, whereas our weighted KNN model reached a peak accuracy of 92.49%. On average, the conventional KNN scored 89.67%, while the weighted KNN averaged 91.89%. The weighted KNN outperformed the conventional KNN.

# 4 Discussion

In this project, a feature weighting technique based on mutual information is introduced and applied to the Wisconsin Breast Cancer (WBC) database. The study compares the performance of weighted models against their original counterparts in terms of classification accuracy.

The results indicate that the weighted approaches match or surpass the performance of the conventional methods across naive Bayes, KNN, and logistic regression classifiers. Particularly noteworthy is the significant improvement observed in accuracy by the weighted KNN model. Being a non-parametric classifier, its potential application extends to diverse datasets, highlighting its usefulness.

However, the weighted versions of naive Bayes and logistic regression models demonstrated limited improvements in accuracy. There is a need for additional experimentation across diverse datasets to explore their potential enhancements.

# 5 References

[1] Tibshirani R. Friedman J. H. Hastie, T. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, Springer., 2st edition, 2009.

[2] P. E. Latham and Y. Roudi. Mutual information. Scholarpedia, 4(1):1658, 2009. revision #186917.

[3] WIlliam Wolberg. Breast cancer wisconsin (original). UCI Machine Learning Repository, 1992. DOI: https://doi.org/10.24432/C5HP4Z.