Milestone1 Report

Machine Bias in Diabetes Dataset

By

647020016-1	Mr. Yordyod Limrostham
655380002-2	Miss Natnarin Phonprapreut
655380003-0	Mr. Jakkrin Srihajak
665380009-9	Mr. Pipat Kumthinkaew

Present

Asst. Prof. Dr. Chitsutha Soomlek

SC348 810 Software Development and Project Management

for Data Science and Artificial Intelligence Semester 1/2566

College of Computing

Khon Kaen University

Semester 1/2566

SC348 810 Software Development and Project Management for Data Science and Artificial Intelligence

---

Milestone#1

Task 1: Study the characteristics of the data and identify the quality issues in the selected dataset.

Dataset details

1. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.
2. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Features definition

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age (years)
- Outcome: Class variable (0 or 1)

Sources:

(a) Original owners: National Institute of Diabetes and Digestive and
Kidney Diseases
(b) Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu)
Research Center, RMI Group Leader
Applied Physics Laboratory
The Johns Hopkins University
Johns Hopkins Road
Laurel, MD 20707
(301) 953-6231
(c) Date received: 9 May 1990

Identify the quality issues in the selected dataset

1. Dataset has some columns that include missing values.

2. Imbalance data Ex. Number of sampling in each classes class 0 (no diabetes) = 500, 1 (has diabetes) = 268

3. Sampling data were collected from specific 1 country, 1 gender and small size of sample in each class.

4. Some features in the dataset have many outlier values.

Task 2: Define the goals and a suitable measure for the quality issues.

Goal and Suitable measure for the quality issues

1. Arrange data to decrease bias in real dataset.
2. Develop model for predicting fairness predicted result.
3. Decrease data bias which is issues from collecting data processes.
    a. Missing data
    b. Data imbalance
    c. Data outliers
4. Analyse model bias.

Suitable measure for the quality issues

1. Countplot : use for checking number of classes in this dataset.
2. Boxplot : use for checking outliers in data.
3. Dataframe : showing and deal with missing values in each features.
4. Confusion matrix : showing model predicted results comparing with actual results.

Task 3: Explore different kinds of machine learning models developed with different modeling techniques. Then, choose the machine learning techniques, implement the models using scikit-learn, and train the models.

1.Naïve Bayes

Naive Bayes is a probabilistic machine learning algorithm commonly used for classification and text analysis tasks, particularly in situations where you need to make predictions based on a set of features. It's based on Bayes' theorem and is considered "naive" because it makes a strong and often unrealistic assumption that the features used to make predictions are conditionally independent given the class label. Despite this simplification, Naive Bayes can work surprisingly well in many real-world applications and is computationally efficient

Bayes' Theorem: The foundation of the Naive Bayes model is Bayes' theorem, which describes the probability of an event based on prior knowledge of conditions related to that event. In a classification context, Bayes' theorem can be expressed as:

$P(y \mid X) = (P(X \mid y) * P(y)) / P(X)$

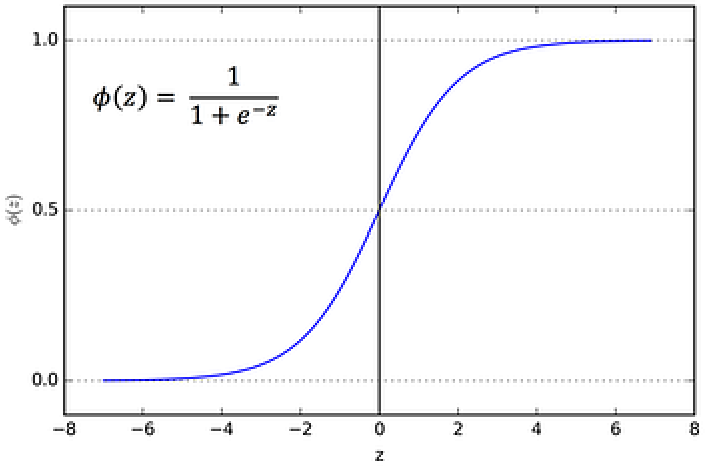
- $P(y \mid X)$: The probability of class y given features X (the posterior probability).
- $P(X \mid y)$: The probability of observing features X given class y (the likelihood).
- $P(y)$: The prior probability of class y.
- $P(X)$: The probability of observing the features X.

Naive Bayes classifiers are simple, fast to train, and can work surprisingly well in various tasks like spam email detection, sentiment analysis, and document categorization. However, their performance may suffer when the independence assumption is significantly violated, or when dealing with high-dimensional data. In such cases, more complex models like decision trees, random forests, or neural networks may be more suitable.

2. Logistic regression

Logistic Regression is a statistical and machine learning model used for binary classification tasks, where the goal is to predict one of two possible outcomes (classes), typically denoted as 0 and 1, based on one or more predictor variables or features. Despite its name, logistic regression is used for classification, not regression.

Sigmoid Function: Logistic regression uses a sigmoid (logistic) function to model the relationship between the input features and the probability of belonging to the positive class (class 1). The sigmoid function is defined as:



Where:

- **z**: The linear combination of input features and their corresponding coefficients (weights).
- **σ(z)**: The output of the sigmoid function, which represents the estimated probability that the input example belongs to the positive class.

Logistic Regression is widely used in various applications such as spam email classification, medical diagnosis, credit scoring, and many other binary classification tasks. It's simple, interpretable, and can serve as a baseline model for more complex classification problems.

3. Linear Discriminant Analysis (lda)

A statistical method used for dimensionality reduction and classification tasks in machine learning and pattern recognition. It is particularly useful when dealing with multivariate data by finding the linear combinations of features that best discriminate between different classes or groups.

LDA is commonly used in fields such as image recognition, biometrics, and medical diagnosis, where the goal is to reduce the dimensionality of data while preserving class-specific information. It's important to note that LDA assumes that the data follows a Gaussian distribution and that the classes have similar covariance matrices. If these assumptions don't hold, the performance of LDA may be compromised.

4. Support Vector Machine (SVM)

Support Vector Machine is a powerful and versatile machine learning algorithm used for both classification and regression tasks. SVM is particularly effective in scenarios where you need to find a hyperplane that best separates data points of different classes in a high-dimensional feature space. SVM aims to maximize the margin between the two classes while minimizing the classification error. Here's an explanation of how SVM works:

In SVM, the goal is to find the hyperplane that best separates data points belonging to different classes. A hyperplane is a decision boundary that maximizes the margin between the two classes. The margin is defined as the distance between the hyperplane and the nearest data points from each class. These nearest data points are called support vectors.

- In a binary classification problem, the hyperplane can be represented as: $w \cdot x + b = 0$
- where w is the weight vector, x is the feature vector, and b is the bias term.

SVMs have several advantages, including their ability to handle high-dimensional data, effectiveness in handling non-linear data with kernel tricks, and good generalization performance. However, they can be sensitive to the choice of hyperparameters, such as the regularization parameter C, and can be computationally expensive for large datasets. Overall, SVMs are a valuable tool in machine learning for various classification and regression tasks.

References

Diabetes dataset : https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset?datasetId=2527538&sortBy=voteCount