**深 圳 大 学 实 验 报 告**

**课程编号：­ 2801000049**

**课程名称：­ 机器学习**

**实验项目名称： 课程项目**

**学院： 电子与信息工程学院**

**专业： 电子信息工程**

**指导教师： 麦晓春**

2022280247

2022280445

2022280307

林茵茵

何雨璇

叶朗钊

**报告人： 学号： 班级： 电信6班**

**实验时间： 2024年 5月28日-2024年6月28日**

**实验报告提交时间： 2024 年 6 月 28 日**

**教务部制**

|  |  |  |
| --- | --- | --- |
| **Classification and forecast of first-hand commercial housing transaction information in Shenzhen**  **林茵茵, 何雨璇,叶朗钊**  **Contributing percentage: 45% + 35% + 20%** | | |
| **Abstract**  **(250-300 words)** | | |
| **1. Introduction**  **(Describe the task background, motivation, the problem you solve, the solution you propose to solve the problem, contributions and novelty)** | | |
| **2. Method (Use at least two pages to illustrate the methodology)**  **2.1 Linear Regression Model**  **2.1.1** **Principle of Linear Regression**  Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to find the best-fitting straight line through the data points that can be used to predict the target variable based on the features.  **2.1.2 Mathematical Formulation**  For a simple linear regression with one feature, the model can be expressed as:  1. is the dependent variable (target).  2. is the independent variable (feature).  3. is the intercept of the line.  4. is the slope of the line (coefficient for the feature).  5. is the error term (residuals).  For multiple linear regression with multiple features, the model extends to:  Where ​ are the independent variables.  **2.2 Ridge Regression Model [1]**  **2.2.1 Principle of Ridge Regression**  Ridge regression is a type of linear regression that includes a regularization term in its cost function to prevent overfitting. This regularization term penalizes large coefficients, effectively shrinking them towards zero. This is particularly useful when dealing with multicollinearity or when the number of features is large compared to the number of observations.  **2.2.2 Mathematical Formulation**  The objective function for ridge regression is given by:  1. is the dependent variable (target).  2. are the independent variables (features).  3. is the regularization parameter (also known as the ridge parameter or shrinkage parameter).  The second term, is the regularization term that penalizes the size of the coefficients. The higher the value of α\alphaα, the stronger the regularization, and vice versa.  **2.3 Logistic Regression Model [2]**  **2.3.1 Principle of Logistic Regression Model**  Logistic regression is a statistical method for analyzing datasets in which there are one or more independent variables that determine an outcome. The outcome is typically binary (0 or 1, True or False, Yes or No). Logistic regression is used to model the probability of a certain class or event, such as pass/fail, win/lose, alive/dead, or healthy/sick.  **2.3.2 Mathematical Formulation**  Logistic regression is based on the logistic function, also known as the sigmoid function, which can be written as:  1. is the logistic function.  2. is the linear combination of input features.  The linear combination of input features zzz is given by:  1. are the independent variables (features).  2. are the coefficients.  The logistic function outputs a probability value between 0 and 1. The decision rule for class prediction is based on a threshold, typically 0.5. If σ(z)≥0.5\sigma(z) \geq 0.5σ(z)≥0.5, the predicted class is 1; otherwise, it is 0.  **2.4 Linear Discrimination Analysis Model [3]**  **2.4.1 Principle of Linear Discrimination Analysis Model**  Linear Discriminant Analysis (LDA) is a classification and dimensionality reduction technique used in machine learning and statistics. It aims to find a linear combination of features that best separate two or more classes of objects or events. LDA is particularly useful when dealing with classification problems where the classes are well-separated.  **2.4.2 Mathematical Formulation**  LDA assumes that different classes generate data based on Gaussian distributions with a shared covariance matrix. The goal is to project the data onto a lower-dimensional space with good class separability.  **2.4.2.1 Within-Class Scatter Matrix**  Measures the scatter (variance) within each class.  Defined as:  where is the number of classes, ​ is the set of samples in class , ​ is a sample in class , and ​ is the mean vector of class .  **2.4.2.2 Between-Class Scatter Matrix**  Measures the scatter between the different class means.  Defined as:  Where  **is the number of samples in class , is the mean vector of class , and is the overall mean vector of the dataset.**  **2.4.2.3 Objective**  Maximize the ratio of the between-class scatter to the within-class scatter:  where www is the projection vector.  **2.4.2.4 Solution**  The optimal projection vectors www are the eigenvectors corresponding to the largest eigenvalues of the matrix ​.  **2.5 Gaussian Naïve Bayes[4]**  **2.5.1 Principle of Gaussian Naïve Bayes**  Gaussian Naive Bayes (GaussianNB) is a classification algorithm based on Bayes' Theorem with an assumption of independence between every pair of features. It is a variant of the Naive Bayes algorithm that is suitable for continuous data, where the likelihood of the features is assumed to be Gaussian (normal distribution).  **2.5.2 Bayes’ Theorem**  Bayes' Theorem describes the probability of an event based on prior knowledge of conditions that might be related to the event. The theorem is expressed as:  1. is the posterior probability of class given the feature vector .  2. is the likelihood, the probability of the feature vector given class .  3. is the prior probability of class .  4. is the marginal likelihood, the total probability of the feature vector .  **2.5.3 Gaussian Distribution**  In Gaussian Naive Bayes, it is assumed that the continuous features associated with each class follow a Gaussian (normal) distribution. The probability density function of a Gaussian distribution is given by:  1. is the feature value.  2. is the mean of the feature for class .  3. is the variance of the feature for class .  **2.6 Multinomial Naïve Bayes[5]**  **2.6.1 Principle of Multinomial Naïve Bayes**  Multinomial Naive Bayes (MultinomialNB) is a variant of the Naive Bayes algorithm tailored for classification with discrete feature vectors. This makes it particularly well-suited for text classification and other applications where data can be represented as frequency counts or other discrete measurements.  **2.6.2 Bayes’ Theorem**  Bayes' Theorem describes the probability of an event based on prior knowledge of conditions that might be related to the event. The theorem is expressed as:  1. is the posterior probability of class given the feature vector .  2. is the likelihood, the probability of the feature vector given class .  3. is the prior probability of class .  4. is the marginal likelihood, the total probability of the feature vector .  **2.6.3 Multinomial Distribution**  In Multinomial Naive Bayes, the likelihood of the features is assumed to follow a multinomial distribution. This is appropriate for discrete data, where each feature represents the count of occurrences of a particular event.  1. is the count of the feature.  2. is the probability of the feature given class .  **2.7 Bernoulli Naïve Bayes[6]**  **2.7.1 Principle of Bernoulli Naïve Bayes**  Bernoulli Naive Bayes (BernoulliNB) is a variant of the Naive Bayes algorithm designed for binary/boolean features. This model is particularly well-suited for tasks where features are binary-valued, such as text classification with binary word occurrence features (e.g., a word is present or not).  **2.7.2 Bayes’ Theorem**  Bayes' Theorem describes the probability of an event based on prior knowledge of conditions that might be related to the event. The theorem is expressed as:  1. is the posterior probability of class given the feature vector .  2. is the likelihood, the probability of the feature vector given class .  3. is the prior probability of class .  4. is the marginal likelihood, the total probability of the feature vector .  **2.7.3 Bernoulli Distribution**  In Bernoulli Naive Bayes, the likelihood of the features is assumed to follow a Bernoulli distribution. This is appropriate for binary data, where each feature represents a binary occurrence (1 if the feature is present, 0 if it is not).  1. is the binary value of the feature (0 to 1).  2. is the probability of the feature being 1 given class .  **2.8 Support Vector Machine[7]**  **2.8.1 Principle of Support Vector Machine**  Support Vector Machines (SVM) are a set of supervised learning methods used for classification, regression, and outlier detection. SVM is particularly known for its ability to perform well in high-dimensional spaces and its effectiveness in cases where the number of dimensions exceeds the number of samples.  **2.8.2 Basic idea**  The core idea of SVM is to find the hyperplane that best divides a dataset into classes. A hyperplane in an n-dimensional space (where n is the number of features) is a flat affine subspace of dimension (n-1). SVM aims to maximize the margin between the two classes, where the margin is defined as the distance between the hyperplane and the nearest data point from either class.  **2.8.3 Support Vectors**  Support vectors are the data points that are closest to the hyperplane. These points are critical in defining the position and orientation of the hyperplane. The decision boundary is thus determined by these support vectors rather than the whole dataset.  **2.8.4 Linear Support Vector Machine**  For linearly separable data, SVM finds a linear hyperplane that separates the data into two classes. Mathematically, given a set of training examples where is a feature vector and ​ is the class label, the goal is to find a hyperplane defined as:  where is the weight vector and is the bias. The optimization problem is to maximize the margin , subject to the constraints for all .  **2.8.5 Non-linear Support Vector Machine**  When the data is not linearly separable, SVM can use kernel functions to project the data into a higher-dimensional space where a linear hyperplane can separate the classes. Common kernels include:  1. Linear Kernel:  2. Polynomial Kernel:  3. Radial Basis Function (RBF) Kernel:  4. Sigmoid Kernel:  **2.9 Decision Tree[8]**  **2.9.1 Principle of Decision Tree**  Decision trees are a popular and powerful machine learning algorithm used for both classification and regression tasks. They work by splitting the data into subsets based on the value of input features, creating a tree-like model of decisions. Each node in the tree represents a feature, each branch represents a decision rule, and each leaf node represents an outcome.  **2.9.2 Basic idea**  The basic idea behind decision trees is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The process of making decisions in a decision tree involves asking a series of questions about the input features and branching out based on the answers until a prediction is made.  **2.9.3 Structure**  1. Root Node: Represents the entire dataset and the first feature to split on.  2. Internal Nodes: Represent the features on which the data is split.  3. Branches: Represent the outcome of the split and connect nodes.  4. Leaf Nodes: Represent the final output or decision (e.g., class label in classification or continuous value in regression).  **2.9.4 Splitting Criteria**  The main task in constructing a decision tree is to determine which feature to split on at each node and what threshold to use for the split. Common criteria for splitting include:  1. Gini Impurity: Measures the impurity of a node (used in classification).  Where is the probability of class in dataset D.  2. Information Gain: Measures the reduction in entropy after a dataset is split on a feature (used in classification).  Where .  3. Mean Square Error (MSE): Measures the average squared difference between the actual and predicted values (used in regression). |
| **3. Experiments and Results**  **3.1 Data Collection, Preprocessing and Analysis**  **3.2 Evaluation Metrics**  **3.3 Experiments**  **3.3 Experimental Results and Analysis** |

|  |
| --- |
| **4. Discussion**  **5. Conclusions**  **References**  [1] What is ridge regression: <https://www.ibm.com/topics/ridge-regression>  [2] What is Logistic Regression? Equation, Assumptions, Types, and Best Practices: <https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/>  [3] What is linear discriminant analysis (LDA): <https://www.ibm.com/topics/linear-discriminant-analysis>  [4] Naïve Bayes: <https://scikit-learn.org/stable/modules/naive_bayes.html>  [5] Multinomial Naïve Bayes: <https://www.geeksforgeeks.org/multinomial-naive-bayes/>  [6] Bernoulli Naïve Bayes: <https://www.geeksforgeeks.org/bernoulli-naive-bayes/>  [7] Support Vector Machine (SVM) Algorithm: <https://www.geeksforgeeks.org/support-vector-machine-algorithm/>  [8] Decision Tree: <https://www.geeksforgeeks.org/decision-tree/> |

深圳大学学生实验报告用纸

|  |
| --- |
| **Appendix**  **Appendix A: xxx**  **Appendix B: xxx** |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  年 月 日 |
| 备注： |

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。

2、教师批改学生实验报告时间应在学生提交实验报告时间后10日内。