Classification (machine learning) Algorithms Data Science Machine Learning Computer Science

Which machine learning algorithms are appropriate for numerical, categorical or both values?

5 Answers

Shehroz Khan, ML Researcher, Postdoc @U of Toronto
Updated Jan 15, 2016 · Author has **1.4k** answers and **4.2m** answer views

For numerical data, choices are too many - starting from basic decision trees, naive bayes, SVM, logistic regression, ensemble methods (bagging, boosting), Random forest, multi-layer perceptron etc.

For categorical data - naive bayes, decision trees and their ensembles including Random forest, Minimum distance classifiers or KNN type with a cost function different than euclidean distance e.g. hamming distance

For 'mixed data', one option is to go with decision trees, other possibilities are naive Bayes where you model numeric attributes by a Gaussian distribution or kernel density estimation or so. You can also employ a minimum distance or KNN based approach; however, the cost function must be able to handle data for both types together. If these approaches don't work then try ensemble techniques. Try bagging with decision trees or else Random Forest that combines bagging and random subspace. With mixed data, choices are limited and you need to be cautious and creative with your choices. (Taken from Shehroz Khan's answer to Which algorithm fits best for categorical and continuous independent variables with categorical response in Machine Learning?)

25.6k views · View 34 Upvoters

Related Questions

More Answers Below

Which algorithm fits best for categorical and continuous independent variables with categorical response in Machine Learning?

How can I use KNN for mixed data (categorical and numerical)?

Which feature selection technique is appropriate for both numerical and categorical values?

What machine learning algorithm should I use when I have 5-6 independent categorical values and 1 dependent continuous variable?

I have 3 numerical and 3 binary features, and the output is a class with two possible values. Which machine learning algorithm fits best this ...

Related Questions

Which algorithm fits best for cate
continuous independent variables
categorical response in Machine L

How can I use KNN for mixed data
and numerical)?

Which feature selection technique
for both numerical and categorica

What machine learning algorithm
when I have 5-6 independent cate
and 1 dependent continuous varia

I have 3 numerical and 3 binary fe
the output is a class with two pos
Which machine learning algorithm
...

Do all known machine learning alg
require numerical inputs (X)?

Alvin Grissom II (グリサム アルビン), Ph.D. Computational Linguistics & Machine Learning, University of Colorado Boulder (2017)

Answered May 15, 2017 · Author has **913** answers and **1.5m** answer views

It depends on the task, but in general, there's nothing stopping you from using numerical values in a categorial model or vice versa.

Various strategies, such as bagging or other means of operationalization, are often used for transforming numerical attributes into categorical ones. Sometimes, a model trained in this way will outperform one using raw numerical values. This is true for both the "labels" and the features. You can, for example, use linear regression for classification, and you can use SVMs to predict numerical ranges created by bagging numerical values.

Intuitively, label bagging makes less sense if you're just trying to minimize the average distance between a predicted numerical value and the actual value.

But, as is often the case in machine learning, if it seems reasonable, it might work, so you can try it and see.

8.7k views · View 1 Upvoter · Answer requested by Mariel Young

Puneet Arora, works at Ecologic Corporation

Answered Sep 15, 2018 · Author has **157** answers and **116.3k** answer views

The answer is simple , you need to do empirical evaluation of many algorithms , fine tune their parameters and find which one the most accurate and lowest false rate .

1. You should typically convert the categorical data into numerical form by using something called code and encoding .
2. Then , you should do some preliminary examination of the dataset with methods such as Descriptive Statistics .
3. Then , some treatment of missing values , erratic values and interpolation to find the trend .
4. Identify the correlation etc
5. Conduct Overlapping/Degree Separation analysis of data variables analysis
6. Then using this knowledge , give default values to the algorithms and start evaluation all the algorithms .
7. There many other factors ... based on which one should choose the machine learning model . Well that can only done after doing some observation on the dataset and after knowing the purpose /objective .

2.1k views · View 2 Upvoters · Not for Reproduction

Prashanth Ravindran, Machine Learning enthusiast

Updated Mar 24, 2018 · Author has **436** answers and **1m** answer views

Based on my experience,

when the input is all numerical (or continuous), logistic regression, random forests, AdaBoost and SVM are good techniques. Try SVM with linear kernel first and then try

with the RBF kernel.

When the input is all categorical, classification trees and random forests are good techniques.

And when the input is mixed numerical and categorical, again classification trees and random forests are good techniques.

Hope it helps.

Related Questions More Answers Below

Do all known machine learning algorithms require numerical inputs (X)?

What are the machine learning algorithms that can be used for health prediction (data set consists of both numerical and binary values)?

What is the best machine learning algorithm to predict numerical data?

What is the k-nearest neighbors algorithm?

How do I use a clustering algorithm on data that has both categorical and numeric value ?

Leon Palafox, Machine Learning an Planetary Science Researcher at University of Arizona

Answered Dec 25, 2015 · Author has **145** answers and **245.6k** answer views

Technically you can use any algorithm for whatever values.

There are plenty of techniques to transform categorical values to numerical ones.
Convert a categorical variable to a numerical variable prior to regression

Now, that said, there are some Natural Language Processing algorithms for which it just does not make much sense to use real data. That is because they count instances of the data, so counting real numbers is just an awful task.

So in conclusion

Continue Reading

Related Questions

Which algorithm fits best for categorical and continuous independent variables with categorical response in Machine Learning?

How can I use KNN for mixed data (categorical and numerical)?

Which feature selection technique is appropriate for both numerical and categorical values?

What machine learning algorithm should I use when I have 5-6 independent categorical values and 1 dependent continuous variable?

I have 3 numerical and 3 binary features, and the output is a class with two possible values. Which machine learning algorithm fits best this ...

Do all known machine learning algorithms require numerical inputs (X)?

What are the machine learning algorithms that can be used for health prediction (data set consists of both numerical and binary values)?

What is the best machine learning algorithm to predict numerical data?

What is the k-nearest neighbors algorithm?

How do I use a clustering algorithm on data that has both categorical and numeric value ?

What is the difference between KNN and K-Means, and what is the better algorithm?

How do I work with numerical categorical features in machine learning?

How do you implement a K-means clustering algorithm on both numerical and categorical data?

How can I apply an SVM for categorical data?

How do I perform a KNN algorithm with a mix of Categorical and Continuous Variables?