Data scientists Documents

Machine learning related questions

Statistics

Probability Question

Model Questions

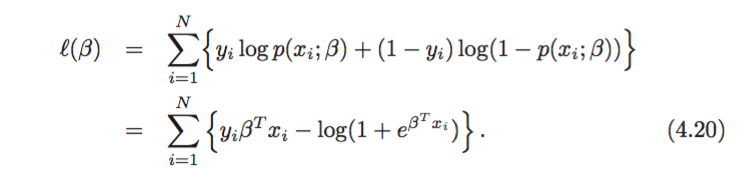
Business Questions

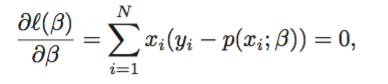
# Machine learning related questions

## Classification:

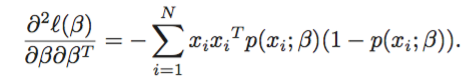
### Logistic regression

Log likelihood function



first derivative

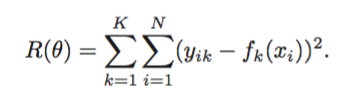
second derivative



y distribution (binomial distribution). Use Newton method calculate the beta.

### Neural Network (classification/regression)

For regression, use sum of square as the cost function



for classification, use cross-entropy as the error function.

ssd:Users:weizhicheng1984:Desktop:Screen Shot 2016-10-03 at 23.54.25.png

Back propagation: in the forward pass, the current weights are fixed and the predicted values are computed. In the backward pass, the errors are computed, and back propagated to give the errors Smi. Both sets of errors are then used to compute the gradients for the updates.

Start values of weights are randomized chosen some values close to 0. It starts with close to linear, and become non-linear.

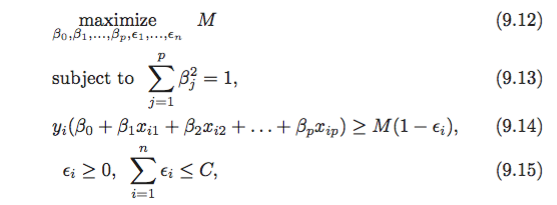
Over fitting: regularization and dropout

Scale the values before put in the model

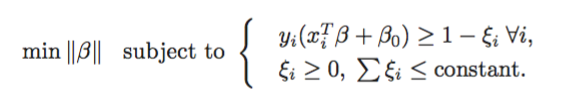
### SVM

Hyperplane is when . If the point is at the one side, then define y =1;if the points lies at the other side then and define y = -1. As a result, if the points can be separated by a hyperplane, all points satisfies

SVM

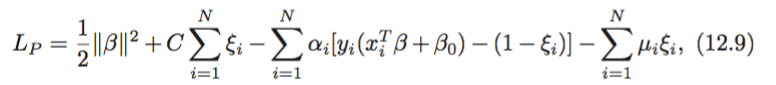


Another way to write this is:



M = 1/||

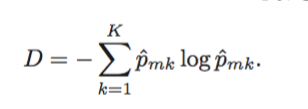
The Lagrange primal function is



Kernel function is to quantify the similarity of a pair of observations.

### Decision tree

For regression, . is the mean response for the training observations within the jth box. For classification, we predict that each observation belong to the most commonly occurring class of training observations in the region. Use Gini index or cross-entropy to quantify the error.



pmk presents the proportion of training observations in the mth region that are from kth class.

Use greedy algorithm to recursive binary splitting the dataset from top to bottom.

### Random forest

Very similar to bagging method, but only use subset of predicators for each small tree. Typically choose m =

### Bayesian network

Don’t understand

### Nearest neighbor classification

Given a query point x0, we find the k training points xr, r= 1, …k, closet in distance to x0, and then classify using the majority vote among the k neighbors. Ties are broken at random.

## Regression:

### Neural Net regression

### Linear regression

Diagnostics plot:

1. Plot residual versus fitted values
2. Normal probability plot on residual
3. Plot residual versus time

Assumption of linear regression model:

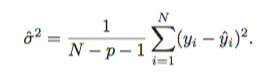
1. Linear relationship
2. The error term is zero mean and normal distribution
3. Y variables is normal distributed and independent
4. Homoscedasticity
5. No or little multicollinearity

Residual sum of squares (RSS)

ssd:Users:weizhicheng1984:Desktop:Screen Shot 2016-09-11 at 22.50.29.png

var(

Residual standard error (RSE) =



SE( vj is the jth diagonal element of (XTX)-1

95 % of confidence interval of beta is

In the multiply regression setting with p predictors, we need to ask whether all of the regression coefficients are zeros. The null hypothesis,

H0 : B1= B2 =B3 = .. Bp = 0

The alternative is at least one of Beta is not zero.

We use F-statistics

F =

Coefficient of determination R2 = 1 – RSS/(total variance)

Model Selection:

1. Exhausted selection
2. Stepwise selection

Model selection criteria: AIC, BIC, PRESS, adjusted R square . AIC is more stable compare with PRESS. Also AIC is easier to calculated

What we do if this happened:

1. Influence data points: calculate cook’s distance for each observation. If the value is larger than 4/n, it is warning sign. Consider using robust regression or remove this point if it is a outlier
2. The relationship is not linear or the residual is not constant, may be do the box Cox power transformation. Use max likelihood to find lambda
3. There is multicollinearity. Use VIF to detect it. Use principal components analysis or lasso

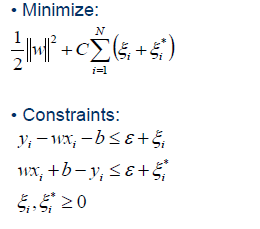
### Ridge regression

Add the L2 loss function

### Lasso regression

Add the L1 loss function

### Support Vector Regression



One of the advantages of ***Support Vector Regression*** as the part of it, is that it can be used to avoid difficulties of using linear functions in the high dimensional feature space and optimization problem is transformed into dual convex quadratic programs. In regression case the loss function is used to penalize errors that are grater than threshold – e. Such loss functions usually lead to the sparse representation of the decision rule, giving significant algorithmic and representational advantages.

### Random forest regression

参考前面的章节

### Partial Least Squaresssd:Users:weizhicheng1984:Desktop:Screen Shot 2016-09-21 at 23.47.13.png

It tries to maximize the variance\*correlation between x and y

## Clustering:

### K-means

Need to run several iterations. The lowest distance is the answer.

### EM

Gaussians mixtures model.

### Mean-shift

### Spectral clustering

Makes no assumption on the shapes of clusters, can handle intertwined spirals. Other clustering method, eg. EM and K means require an iterative process to find local minima and multiple restarts.

Three steps:

1.      Construct a similarity graph (e.g. KNN graph) for all the data points

2.      Embed data points in a low-dimensional space (spectral embedding), in which the clusters are more obvious, with the use of the eigenvectors of the graph Laplacian (degree matrix – adjacency = laplcaian ).

3.      A classical clustering algorithm (e.g. k-means) is applied to partition the embedding.

<https://www.youtube.com/watch?v=vrnfB81CMR0>

argmin f’Lf

f is the solution. L is laplcian

Most powerful clustering method.

### Hierarchical clustering

Do not require specify the cluster number, but it require user to specify a measure of dissimilarity between groups of the observations. There are methods: agglometrative clustering (bottom up) and divisive clustering (top down).

The agglometrative clustering begin with every oberavation representing a singleton cluster. At each of the N-1 steps the closet two clusters are merged into a single cluster, producing one less cluster at the next higher level. Three different way to define the dissimilarity between two clusters: single linkage (least dissimilar two), complete linkage (most dissimilar two) and group average.

Divisive clustering algorithms begin with the entire data set as a single cluster, and recursively divide one of the existing clusters into two daughter clusters at each iteration in a top-down fashion.

## Dimension Reduction:

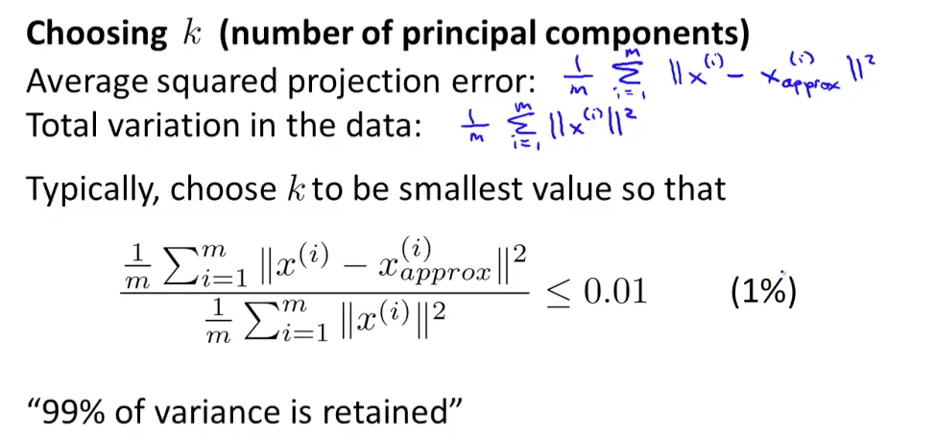
### PCA

Need to do the mean normalize and optionally feature scaling

Calculate covariance matrix:

[U, S, V] = svd(sigma)

Z = UX (first k column of U)



### Independent component Analysis

The independent component analysis model is similar to before:

X =AS, but si is independent.

### Linear discriminant analysis

It is assuming the data is Gaussian distribution. It tries to maximize the separation of known categories . But this is a supervised learning.

LDA的原理是，将带上标签的数据（点），通过投影的方法，投影到维度更低的空间中，使得投影后的点，会形成按类别区分，一簇一簇的情况，相同类别的点，将会在投影后的空间中更接近。

<http://www.cnblogs.com/LeftNotEasy/archive/2011/01/08/lda-and-pca-machine-learning.html>

## SVM formulation. What is the decision boundary?

Decision boundary is the hyper plane

## Logistic regression. Write down the formula for logistic regression. How to determine the coefficients given the data?

Get the first derivative and second derivative. Use the newton’s method to get the beta.

## Regularization

Lasso and rigid regression. It is used to prevent the over fitting.

## Cost function of neural network

For regression, it is sum of error. For classification, it is the cross entropy.

## What is the difference between a generative and discriminative algorithm

## Relationship between kernel trick and dimension augmentation

## What is PCA projection and why it can be solved by SVD

<http://stats.stackexchange.com/questions/130721/what-norm-of-the-reconstruction-error-is-minimized-by-the-low-rank-approximation>

## Bag of Words (BoW) feature

## Nonlinear dimension reduction (Isomap, LLE)

## Supervised methods for dimension reduction

Linear discriminant analysis

## What is naive Bayes.

## What’s the formula for Naive Bayesian classifier? What’s the assumption in the formula? What kind of data is Naive Bayesian good at? What is not?

## Stochastic gradient / gradient descent

## Assumptions of simple linear regression

## Variance and Bias (a very popular question, watch Andrew’s class)

## Practices: When to collect more data / use more features / etc. (watch Andrew’s class)

## How to extract features of shoes

## What’s the difference between classification and regression?

## During linear regression, when using each attribute (dimension) independently to predict the target value, you get a positive weight for each attribute. However, when you combine all attributes to predict, you get some large negative weights, why? How to solve it?

## Cross Validation

## Reservoir sampling

## Explain the difference among decision tree, bagging and random forest

## What is collaborative filtering

## What is quantile regression

## How do you evaluate regression?

For example, in this particular case:

item click-through-rate predicted rate

1 0.04 0.06

2 0.68 0.78

3 0.27 0.19

4 0.52 0.57

How to evaluate classification result? What if the results are in probability mode? If I want to build a classifier, but the data is very unbalanced. I have a few positive samples but a lot of negative samples. What should I do?

# Statistics

## 44.3% vs 47.2% is it significant?

Need do t-test.

## Given a lot of data, I want to random sample 1% of them. How to do it efficiently?

## 1-sample proportion z-test

## Test equal variance among multiple groups

## What is the definition of odds? how to compute CI for odds?

## definition of standard error

## A/B testing, determine sample size. what will happen to p-value if sample sizes increase?

## hypothesis testing

# Probability Question

## Given a coin, how to pick 1 person from 3 persons with equal probability.

One person chooses HH, second person chooses TT, and third person chooses HT. If it is TH, start over again.

## How to compute the average of a data stream (very easy, different from moving average)

X = (X\*n+a)/(n+1) a: the new number X: average

## Given a coin you don’t know it’s fair or unfair. Throw it 6 times and get 1 tail and 5 head. Determine whether it’s fair or not. What’s your confidence value?

Probability = 6\*0.5^6 . It is very unlikely to be a fair coin

## A field with unknown number of rabbits. Catch 100 rabbits and put a label on each of them. A few days later, catch 300 rabbits and found 60 with labels. Estimate how many rabbits are there?

500

# Model Question

## Given Amazon data, how to predict which users are going to be top shoppers in this holiday season.

## Discuss how to predict the price of a hotel given data from previous years

## What is the real distribution of click-through rate of items? If you want to build a predictor/classifier for this data, how do you do it? How do you divide the data?

## You have a stream of data coming in, in the format as the following:

item\_id, views, clicks, time

1            100     10         2013-11-28

1            1000   350       2013-11-29

1            200     14         2013-11-30

2            127     13         2013-12-1

…

The same id are consecutive.

Click through rate = clicks / views.

On every day, I want to output the item id when its click through rate is larger than a given threshold.

For example, at day 1, item 1’s rate is 10/100=10%, day2, its (10+350)/(100+1000)=0.32. day3 it is (10+350+14)/(100+1000+200)=0.28.

If my threshold is 0.3, then at day 1, I don’t output. On day2 I output. Onday3, I don’t output.

## We want to add a new feature to our product. How to determine if people like it?

A/B testing. How to do A/B testing? How many ways? pros and cons?

## Design a function to calculate people’s interest to a place against the distance to the place.

## Given yelp data, how to find top 10 restaurants in America?

## Design a job search ranking algorithm on glassdoor

## How to identify review spam?

## How to predict the age of a person given everyone’s phone call

# Business Questions

## How to encourage people to write more reviews on Yelp? How to determine who are likely to write reviews? How to increase the registration rate of Yelp? What features to add for a better Yelp app? We are expanding to other countries. Which country we should enter first?

# Database & programming:

## We have a database of <product\_id, name, description, price>. When user inputs a product name, how to return results fast?

## If user gives a budget value, how to find the most expensive product under budget? Assume the data fits in memory. What data structure, or algorithm you use to find the product quickly? Write the program for it.

## Given a large file that we don’t know how many lines are there. It doesn’t fit into memory. We want to sample K lines from the file uniformly. Write a program for it.

## Write a function to compute sqrt(X)

## Write a function to compute pow(x, n)

## In a city where there are only vertical and horizontal streets. There are people on the cross point. These people want to meet. Please find a cross point to minimize the cost for all the people to move.

## Youtube has video play log like this:

Video ID, time

vid1        t1

vid2        t2

...           ...

The log is super large.

Find out the top 10 played videos on youtube in a given week.

## Write a program to copy a graph

Leetcode 原题

# GOOGLE:

1. Integral image . 其实就是生成一个2D矩阵，里面的每一个点

(i,j)都记录了从这个点到(0,0)所形成的矩形的元素之和。

2. 一条线段长度L，雨滴直径D，雨滴从不同位置等概率下降到线段上，模拟一下这个情况，求出多少雨滴可以把线段覆盖完全

3. 一个file里面有很多alarm，每个alarm有三个数值：起始时间、终止时间、优先度。把那些从没成为过最高优先度的alarm去除。

4. rotate array by k steps（leetcode），要最优解：reverse不能用

5. （1）俄罗斯方块，求出正在下落的物体和底部的最短距离。

（2）一幅图有一系列二围的点，判断此图是否对称

6. 一堆data, Bernoulli distribution, 40% = 1, 60% = 0, design test to see if 40% is significantly less than expected?

我说用chi-square, 因为是non-parametric, expected frequency = 0.5, 他又问chi-square的公式。

7. 另一堆data, 70% = 0， design a test to see if the two samples （和前面一道题的60%） are different, 我说用two sample t test, 这个对不？

8. 3 columns of data, two continuous variable, one boolean variable, design experiment to see relationship between response and variables

我说用 linear regression, 他问要一步步怎么做要说清楚，我说先 center and standardize continuous variable, 然后看 Boolean 的frequency，是不是 skewed.

还要要看有没有 multicollinearity issue 和 heteroskedasticity, 所有的 assumption 满足了才能做 regression， 他又问有collinearity怎么办，不是 perfect collinearity, 我说用 ridge regression.

8. additional links:

<http://www.datasciencecentral.com/profiles/blogs/66-job-interview-questions-for-data-scientists>

<http://alexbraunstein.com/2011/08/09/hire-data-scientist-statistician/>

<http://alexbraunstein.com/2011/08/09/hire-data-scientist-statistician/>