TODO

Ruichen Wang

February 25, 2019

Some Basics

C4.5, ID3, CART, K-means, SVM, Kernel function, SMO, EM, Naive bayes, Variational inference, PageRank, Adaboost, KNN, Data cleaning, Linear regression, Logistic regression, LDA, PCA, Random Forest, Bagging & stacking, Softmax, GBDT, Xgboost, LightGBM, PCA, Max-entorpy, Gradient diminishing, FM, LFM, MF, SVD, SVD++, Simhash, Max likelihood, ALS, L1 norm, L2 norm, Discriminate & Generative model, Entorpy, Cross-entropy, KL divergence, SGD, BGD, MBGD, Adam, Newton's method, Unbiased estimator, F1 score, Recall, Prercision, AUC, ROC, Cross validation, Bias-Variance tradeoff, Loss function, Overfitting & underfitting, DBSCAN, TF-IDF, Text rank, Word2vec, Glove, Fast-text, Collaborative Filtering, User-cf, Item-cf, Dropout, Batch norm, Maxpooling, Avg pooling, Global avg pooling, Respective field, LR normalization, Online learning, FTRL, FOBOS, RDA, Tree row & column sampling, ANN, Tree normalization, RNN, LSTM, GRU, CRF, RNN & CNN, KDtree, Resnet, Seq2seq, Deep wise CNN, 1X1 kernel, SVM multi-classification, Attention, Conv complexity, Pearson correlation, Gibbs sampling, t-SNE.

---

1 ID3, C4.5, CART difference?

What is Decision Tree (DT)?

DTs are non parametric superivised learning method, can be used for classification and regression. DTs need little data preparation. No need NOrmalization. White box model. Easily overfitting, sensitive to variation, because may build complete different trees, can be mitigated by using ensemble. Based on greedy opt to solve NP-complete problem, may reach local optimal. Need balance dataset.

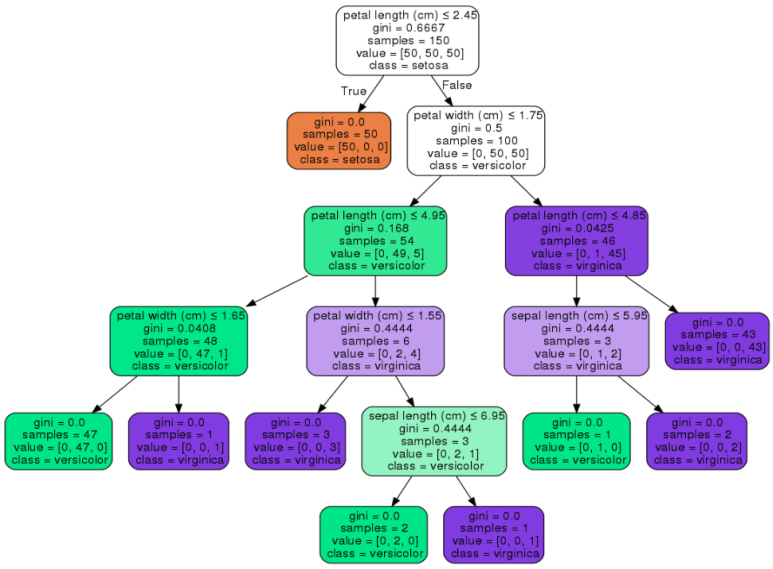

Figure 1: A decision tree using gini.

Pratical Tips:

- DT tend overfit if using a large number of features, Using a ratio to sample the features.
- Consider perform feature dimensionality reduction (PCA, Feature selection).
- Control max depth, control min samples per leaf to prevent overfitting

ID3 (info gain)
Entropy $H(S)$ is a measure of uncertainty of data.

$$H(S) = \sum_{x \in S} -p(x)logp(x)$$

Information Gain $IG(S)$ is a measure of differenece in entropy before and after S splited on attribute A.

$$IG(S, A) = H(S) - \sum_{t \in A} p(t)H(t) = H(S) - H(S|A)$$

C4.5 (gain ratio) C4.5 is the successor to ID3. Can handle continuous attribute.

$$GainRatio(S, A) = \frac{IG(S, A)}{IV(A)}$$

CART(Gini impurity) CART using binary split, Split based on one variable. Pruning mainly has two types : pre and post. CART use CCP (cost complexity pruning), which is defined by $\frac{errorrate}{treecomplexity}$

$$Gini(A,S) = 1 - \sum_{t \in A} p_t^2$$

2 Why L1 regulation generates sparsity? L2 regulation cause blur?

Firstly, why do we want the result matrix to be sparse?

Consider 1 million dimension, calculate the inner product between w and x need a lot of computation. If the w can be sparse, the inner product will only be performed on the non-zero columns.

Or consider another situation, in some scenario, there are free data and many features, which is often called as **'small n, large p problem'**. If $n \ll p$, then our model will be very complex, our w will be a singular matrix ($|w| = 0$). In other words, **overfitting**.

One way to control overfitting is adding a regularization term to the loss function. Rigde ($l_2 norm$) and LASSO ($l_1 norm$) regression are two very common regression ways.

$$J(w) = Loss(x) + \lambda ||w||_2^2$$
$$J(w) = Loss(x) + \lambda ||w||_1$$

Assume we use loss using MSE, the target function can also be denoted as :

$$\min_w \frac{1}{n}||y - Xw||^2 \quad s.t. \lambda ||w||_2^2 \leq C$$
$$\min_w \frac{1}{n}||y - Xw||^2 \quad s.t. \lambda ||w||_1 \leq C$$

Back to the problem, intuitivly, the target loss will alway intersect at the coordinate axis when using l1 norm. Imaging high dimension situation, the angles will certainly more likely to be intersected, while the ball will not.

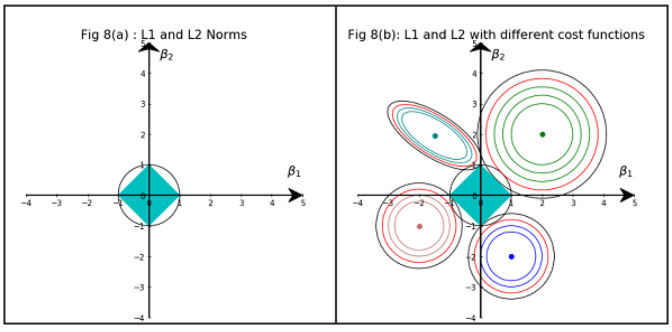

Figure 2: L1 and L2 norm.

For more math proof, see *http://freemind.pluskid.org/machine-learning/sparsity-and-some-basics-of-l1-regularization*

Why L2 regulation cause blur? In generative models, eg.VAE, L2 norm / L2 loss / MSE tend to yield blurry images. We try to explain this in probabilistic settings. In Gaussian distribution, it defines as :

$$p(x|\mu,\sigma^2) = \frac{1}{Z} exp\left(-\frac{||\mu - x||^2}{2\sigma^2}\right)$$

$$logp(x|\mu,\sigma^2) \propto exp\left(-\frac{1}{2}||x_\mu - x||^2\right)$$

As we can see, minimizing MSE, is same as maximizing the log likelihood of gaussian, we make the assumption that our x comes from a gaussian.

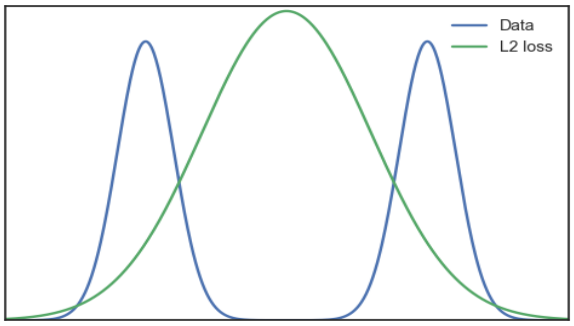

Figure 3: Gaussian and multinomial.

In reality, there often many hidden variables (multinomial) controls the x. A simple example, we have white and black dogs as dataset x. maximizing the likelihood will blur the two and generate gray dogs.

3 One-hot encoding for gbdt?

It is often the case that we have continuous and categorical features. One-hot encoding may produce very sparse variables. Tree based algorithm tries to increase information gain on the data it's splitting at any given level. If the data is very sparse, the one-hot encoded feature may be ignored as they are far too sparse.

Then how to using categorical features in trees?

In xgboost [1], it treat every input as numerical, It may be helpful is the categories as small. Otherwise, it may downgrade the performance.

In lightGBM, it said that a tree built on one-hot features tends to be unbalanced and needs to grow very deep to achieve good accuracy. It's automaticly

provide a optimal solution $O(k * logk)$ to find the best optimal partition. This often performs better than one-hot encoding.

For more info, see lightgbm api.

4 xgboost, lightGBM, gbdt, GBM?

Gradient tree boosting is also known as gradient boosting machine (GBM) or gradient boosted regression tree (GBRT).

xgboost using taylor expansion to approximate the objective function. Regularization is defined as number of leafs plus l2 norm of leaf values.

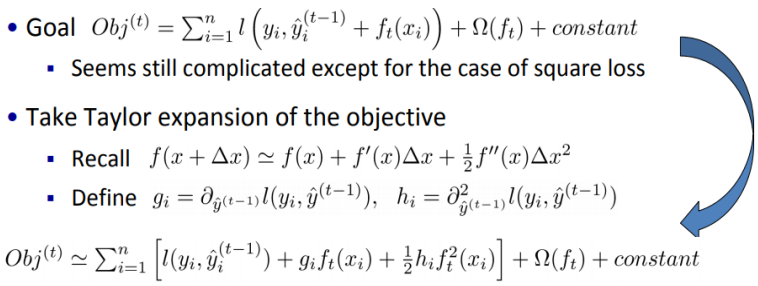

Figure 4:

for detail see pdf in ml papers basis.

lightgbm choose the leaf with max delta loss to grow(leaf wise, best first), while most tree grow by level wise. Using max depth to control overfitting.

5 GD, BGD, SGD, momentum, adagrad, adam

GD refers to gradient descent, Using to negative gradient to optimize the ojbective function. Batch gradient descent (BGD) computes the gradient on the whole dataset, which if often slow and intractable. Can not be applied to online model which data is on the fly.

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta)$$

BGD guaranteed to converge to the global minimum for convex problem.

SGD perform an update for each training example. Usually slowly decrease the learning rate.

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta; x^i; y^i)$$

Mini-batch gradient descent perform an update for every mini batch of n train examples.

Momentum adagrad,adam, rmsprop etc are all opt strategy.

6 EM Algorithm

In the topic of clustering, it can be classified as hard and soft clustering. EM is a soft clustering method which is a generative model which trys calculate parameters of probability distribution. EM algorithm is used for maximum likelihood estimates of parameters. In bayesian statistics, it is often used to obtain the mode of the posterior marginal distributions of parameters.

E-step computes the expected value of $l(\theta; X, Z)$ given the observed data X, and current estimate θ_{old}.

$$Q(Z; \theta_{old}) = P(Z|X, \theta_{old})$$

M-step maximizing over θ the expectation computed.

$$\begin{aligned}\theta_{new} :&= \arg\max_{\theta} -KL(Q(Z)||P(Z|X,\theta)) \\ &= \arg\max_{\theta} \sum_i \sum_{z^i} Q_i(z^i) log \frac{p(x^i, z^i; \theta)}{Q_i(z^i)}\end{aligned}$$

In variational inference, we know that

$$logp(x) = KL(q(z)||p(z|x)) + (E_q[logp(z,x)] - E_q[logq(z)])$$

The difference of EM and VB is the kind of results they provide: **EM is just a point, VB is a distribution**. However, they also have similarities. EM and VB can both be interpreted as minimizing some sort of distance between the true value and our estimate, which is the Kullback-Leibler divergence.
For example:

When the Gaussian model involves latent variables and parameters only, Expectation Maximization is enough to solve the model.

If, on top of the latent variables, the parameters becomes random with prior distributions, the Variational Inference (or Variational Bayes) method is used.

References

[1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 785–794. ACM, 2016.