# Bayesian decision theory

## Regression

Let be a random input vector and is a real-valued random output variable, with joint probability . We seek a function which predicts given values of the input . We assume and are both continuous random variables (regression problem).

In Bayesian decision theory, the function is also called *action* or *strategy* or *decision function*. For a certain input , the *loss* caused by taking action is



where is the (unknown) true output value and is the *loss* function. For a particular observation , the *expected loss* for a given input is called the *conditional risk*, defined by



Thus, for a particular input , we can minimize the expected loss by choosing action that minimizes the conditional risk:



For every possible input , the function denotes a decision rule and the **overall risk** with respect to and is



which is the expected loss in the whole space. Thus, we seek a decision rule to minimize the overall risk: . Equation is a double integral and according to Fubini’s theorem, it can be rewritten into



To minimize the overall risk , it suffices to minimize it pointwise, i.e., at each possible value of X. Then, we get



Thus, we see that if we can minimize the conditional risk for each , then the overall risk must also be minimized. For regression, a population loss function is the squared-error one, that is



With this squared error loss, equation is



We can also take the loss function, , which is more robust than the one. In this case, we have



Therefore, the optimal least squares predicator is the conditional expectation. For the loss, the optimal predicator is the conditional median.

## Classification

For a *K*-class classification problem, suppose the classes are . For a new input , the true class is and our decision rule assigns to class . The loss function is



The conditional risk (expected loss) for this specific input is



because the true class is actually unknown (a discrete random variable) for the given observation.

Similarly, the overall risk is



Furthermore, the overall risk can be expressed in expectation form by



To minimize the overall risk, it suffices to minimize it pointwise



For 0-1 loss, the loss function is



Then, the decision rule in is



while



Therefore, to minimize the 0-1 loss, our decision function chooses the class for which the conditional probability is maximized



# Empirical risk

The fundamental problem with Bayes Decision Theory is that we usually don’t know the joint distribution of the input and output variables . Instead, we only have a set of training data . Therefore, instead of the overall risk (population expectation of the loss function), we define the **empirical risk**



which is also usually named as the *cost function*. A fundamental assumption of BDT and ML is that the observed data *D* consists of independent identically distributed i.i.d samples from an (unknown) distribution . Then, as , we have .

This suggests several strategies to apply Bayesian decision theory in practical data, including

* First use the training data D to learn the distribution ( and ); then apply the decision rule defined in , or . (This is the classic statistics strategy: generative model.)
* The second strategy is the discriminative approach. Estimate the best decision rule directly from the empirical risk in . Motivation: why estimate the probabilities when you only care about the decision? (This is the classic Machine Learning Strategy.)
* Learn the posterior distribution directly.

# Boosting

As mentioned above, to minimize the overall loss, we can minimize the pointwise conditional risk



We seek an additional model . Currently, we have and we want to find another function such that can minimize the conditional risk in , that is



Generally, it is difficult to solve the above optimization problem. However, we can follow the steepest descent algorithm but in the function space instead of data space. The gradient of with respect to is



Therefore, the new function component should be parallel to the negative gradient at . That is,



with a positive scalar and



From the definition of the conditional risk in , we further have



The step size is solved according to the steepest descent principle



In practice, we only have finite training data instead of the joint probability , which means the conditional risk for the input , i.e., cannot be accurately estimated. Therefore, instead of the overall risk, we try to minimize the empirical risk (the cost function)



At the finite data points, the gradient at analogous to is



However, this gradient is only defined at the data points. To generalize it to other possible , we can choose another model to fit these negative gradient data points in a least square sense. Supposing is a parametric model with parameters , then the parameters are solved in least squares by

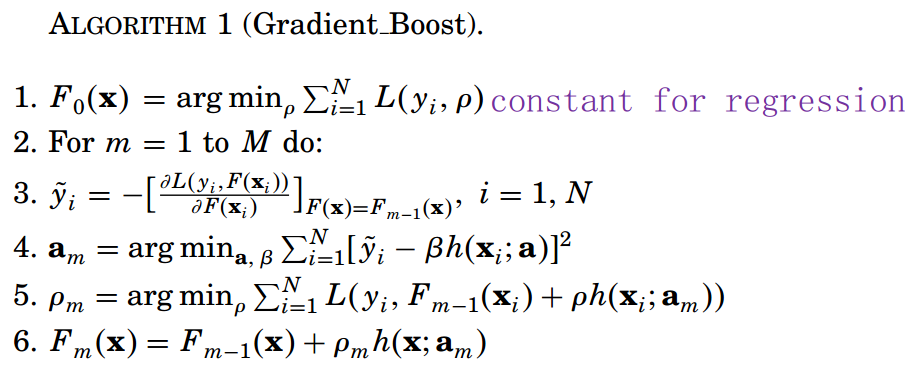


After we get , similar to , the optimal line search is performed



and the update law is





### Summary:

* Because only the direction of the negative gradient matters, the above can be ignored.
* Compute the gradient  at each to be
* Fit (usually with a shallow decision tree)
* Due to the shrinkage for regularization, there is in fact no need to perform line search for . Simply use a shrinkage factor and update .

# Reference

1. The elements of statistical learning.

2. Pattern recognition and machine learning.

3. Lecture Note 2 - Bayesian Decision Theory. Prof. Alan Yuille. <http://www.cs.jhu.edu/~ayuille/courses/Stat161-261-Spring14/Stat_161_261_2014.html>