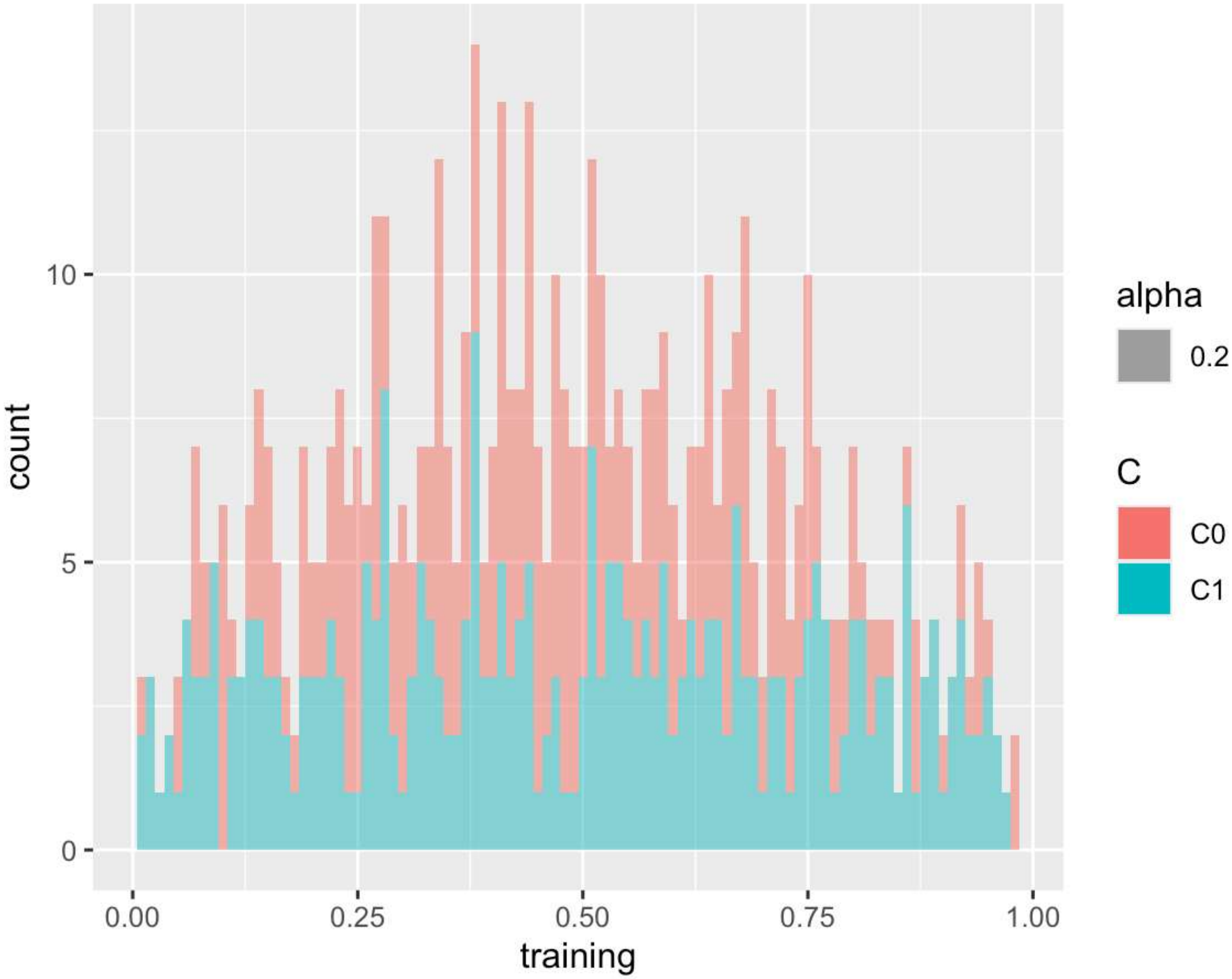
Solutions to Classification Homework

This homework will be a great practice midterm, as it covers the important concepts from the first half of the class. Also, it allows you to begin with training data, and produce (the best) classifiers of all types based on that data. Further, you will be able to estimate the performance of all these classifiers for single measurement, as well as many more measurements.

This homework (and midterm) will be eased by the used of the Supervised ML scripts I have provided. They are written in R (free) using RStudio (free), but can easily be translated to Matlab and Python. The choice is up to you!

1.) (10 points) Please load the linked csv files, which contain measurements for class 0 and class 1. Produce a single plot showing the histograms for class 0 and class 1 measurements. Label the horizontal axis "training" and the vertical axis "count". Referring to example distributions on Wikipedia, which family of distributions do the class-0 measurements belong to? Class 1? Provide reasoning for your answers. Also, please answer the following: how many bins does each histogram have? How many "knobs" need to be adjusted to tune the histogram to the measurements?



The histograms are both bounded between 0 and 1, so the beta family seems to be a great family to use. For B bins, a histogram has B-1 "knobs" to adjust.

2.) (10 points) The following applies to each class separately. Refer to the families of distributions on Wikipedia to view the probability density functions. Find the family of distributions which "looks" like the histogram for class i. Find the link for "statistical inference" and "parameter settings" for that family of distribution. How many "knobs" need to be adjusted for this family? Compare this answer to those given in 1.) and justify the approach for avoiding tuning histograms.

For the beta family on the measurement interval [0,1], there are 2 parameters to estimate. This is far less then B-1 (at least for the histograms here), so we should avoid using histograms directly in our further work.

3.) (10 points) Use the method on Wikipedia to estimate the parameters for each class. Provide the results of your calculations for each class, including the chosen family. Also, prepare a SINGLE plot of the LOGARITHMS of each probability density function. These are called the log-likelihood functions for your classifier and together form the optimal processor for each measurement for classification. On the same plot, also provide the DIFFERENCE of the log-likelihood function for class 1 minus that for class 0. Label the horizontal axis "measurement" and the vertical axis "log-likelihood functions". Provide a legend and coloring to clearly distinguish the three curves.

Wikipedia recommends the method of moment estimators for the beta parameters "a" and "b". This must be done for each class separately, using estimates of the mean (x_bar) and variance (v_bar) below.

$$\hat{\alpha} = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1\right),$$

$$\hat{\beta} = (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1\right),$$

Here, alpha is "a" and beta is "b". The results of the estimates are as follows:

class 0

esta0 = 1.98

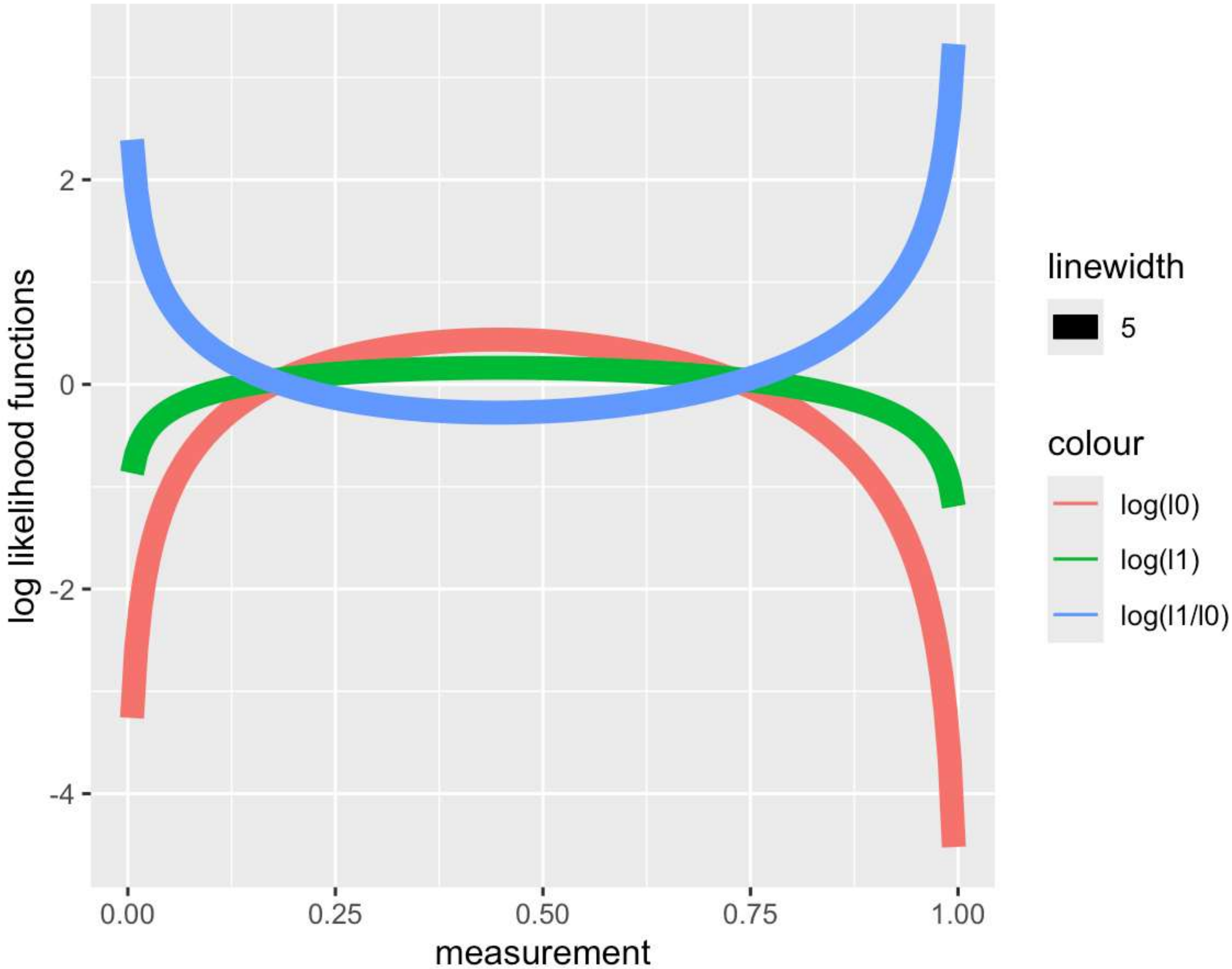
estb0 = 2.22

class 1

esta1 = 1.27

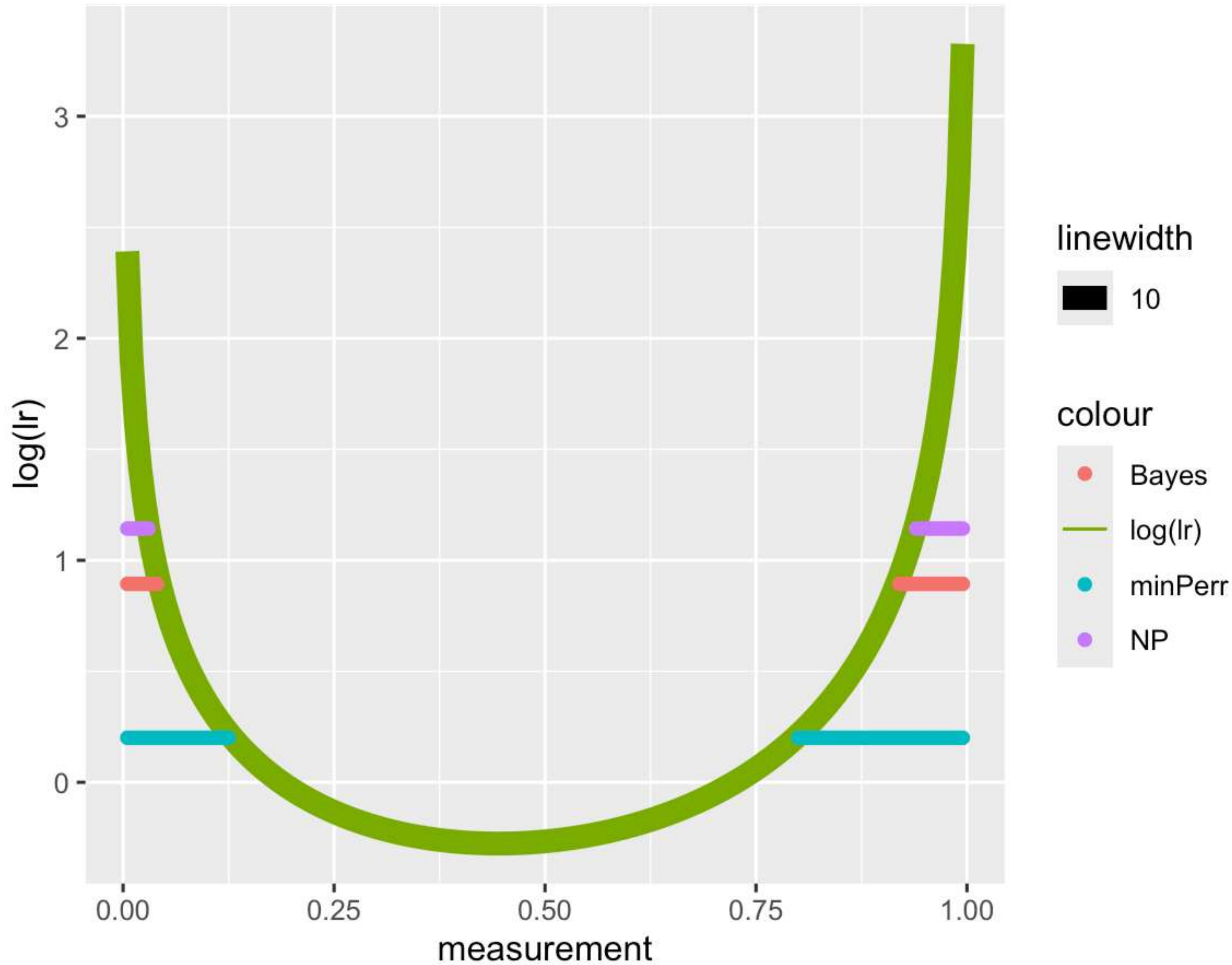
estb1 = 1.33

For these values the requested plot is



4.) (10 points) Please select the thresholds for three tests: the minimum risk (Bayes) classifier, the minimum probability of error classifier, and the Neyman-Pearson classifier. For that, we need some information about the classes and costs for errors. Let the prior probability of class 0 be 0.55, the cost of wrongly saying class 1 be $2, the cost of wrongly saying class 0 be $1, and the maximum false alarm (saying class 1 when you are wrong) probability be 1/100. For this information, please find the thresholds to compare to the LOG LIKELIHOOD RATIO. Produce a SINGLE plot with the log likelihood ratio. Further, mark the decision region for class 1 for each test AT THE HEIGHT OF THE LOG THRESHOLD. This will make the SINGLE plot less cluttered.

The requested plot is



5.) (10 points) Provide a numerical calculation for: the minimum risk of the Bayes detector, the minimum probability of error of the minPerr detector, and the power (the probability of saying class 1 when you are correct) of the NP detector. These values all for a single measurement and may be disappointing. Stay tuned!

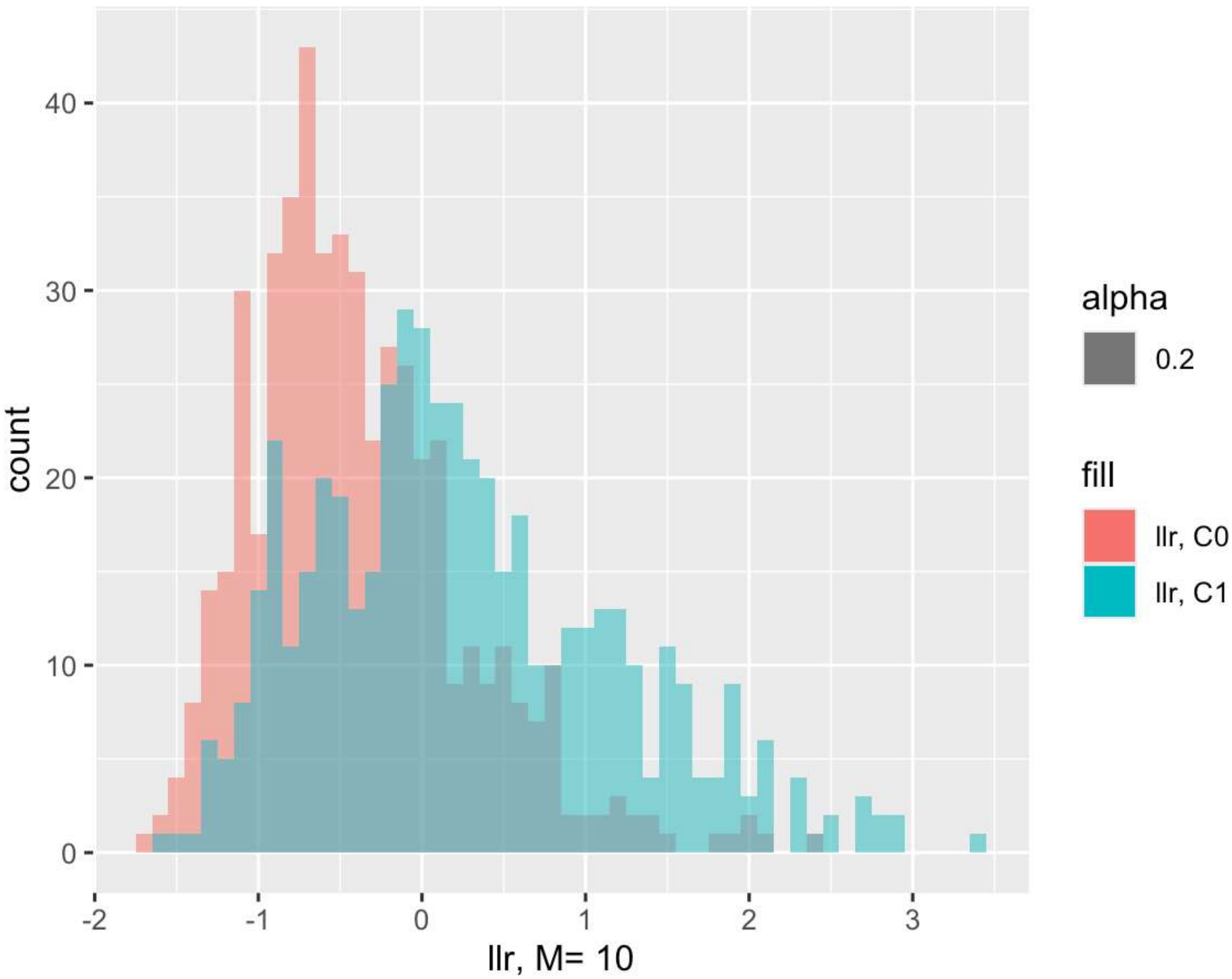
For a single-sample detector, the requested values are:

minimum risk = $ 0.44

minimum probability of error = 0.41 (almost 0.45!)

Power = 0.05

6.) (20 points) To improve performance, we consider more "looks" (measurements) before we make a decision. Consider the log likelihood ratio when there are multiple, independent measurements, each with the same distributions you have found. In that case, the optimal processor takes each measurement, evaluates the log-likelihood ratio found in 3.), and sums the results. Further, we know (Central Limit Theorem) that the sum of the log-likelihood ratios "looks" Gaussian. This helps us do two things: set thresholds, and evaluates performance. Consider the experiment of generating M=10 independent measurements from class 0, passing each through the log-likelihood ratio in 3., and summing the results. This produces one instance of the experimental output. Do this many times to display the histogram of this log-likelihood ratio for class 0 on a graph. ON THE SAME GRAPH, provide the similar histogram of this log-likelihood ratio under class 1. Make the argument that these histograms are bell-shaped curves, and that the log-likelihood ratio is approximately Gaussian for M=10 or more samples.



7.) (10 points) Numerically find the single-sample (M=1) means and variances of the log-likelihood ratio under each class. Use the given data for them, passing them through the log-likelihood ratio first. Multiply those moments by M to find the M-sample mean m_i and variance v_i of the M-sample log-likelihood ratio under class i. For any threshold t, then the probability P[llr >t | class i] is given by the Gaussian distribution with mean m_i and variance v_i.

M=1 sample log-likelihood ratio has mean and variance under class 0 as

m0 = -0.05

var0 = 0.126

and under class 1

m1 = 0.076

var1 = 0.198

You can see these in the histograms of 6.). Multiply each by M for the M-sample case.

8.). (20 points) Set M=10, 50, 100. Each time, find: the risk of the minimum risk detector, the probability of error of the minimum probability of error detector, and the power of the NP test. Using a table with 4 rows: M=1 (from 5.), M=10, M=50, and M=100. Set the columns as: minimum risk, minimum error rate, and power. Using this information to show that multiple, independent "looks" makes decisions as accurate as you want!

M	min risk	min error rate	power
1	0.44	0.41	0.05
10	0.35	0.30	0.17
50	0.17	0.13	0.57
100	0.07	0.06	0.85