Feature Engineering

Chapter 12

Stephen Kimel

Naïve Bayes

Naïve Bayes Models

- Naïve = assumption that predictors are independent

$$Pr[Class|Predictors] = \frac{Pr[Class] \times Pr[Predictors|Class]}{Pr[Predictors]}$$
$$= \frac{Prior \times Likelihood}{Evidence}$$

Normalization function

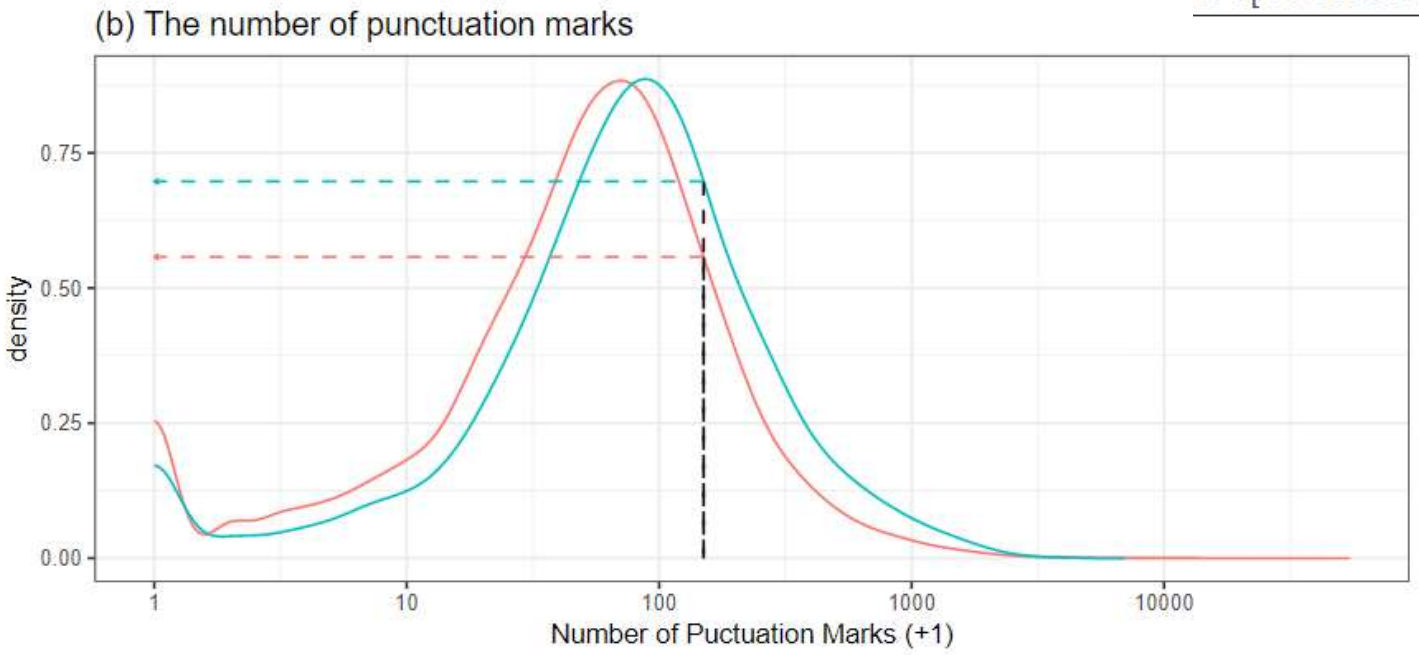
In English:

> Given our predictor data, what is the probability of each class?

Categorical Variable

Continuous Variable

$$Pr[Predictors|Class]$$



Calculations

$$Pr[Class|Predictors] = \frac{Pr[Class] \times Pr[Predictors|Class]}{Pr[Predictors]}$$
$$= \frac{Prior \times Likelihood}{Evidence}$$

Table 12.1: Values used in the naive Bayes model computations.

	Predictor Values								
Class	Religion		Punctuation		Likelihood		Prior		Posterior
STEM	0.213	*	0.697	=	0.148	*	0.185	.027	0.37
other	0.103	*	0.558	=	0.057	*	0.815	.046	0.63

Posterior = $Prior_i * Likelihood_i$ / $\sum_{i=1}^{n} \text{Prior}_i * \text{Likelihoodi}$

Simulated Annealing

Simulated Annealing

Find Most Predictive Set of Features

1. Random Subset of Features
2. Chose Number of Iterations
3. Build Model
4. Calculate Performance
5. Randomly Include/Exclude 1-5% of Features
6. If the Model Improved Keep New Model and Repeat Process
7. If Model Did Not Improve…

Model Did Not Improve

$$Pr[accept] = \exp\left[-\frac{i}{c}\left(\frac{old - new}{old}\right)\right]$$

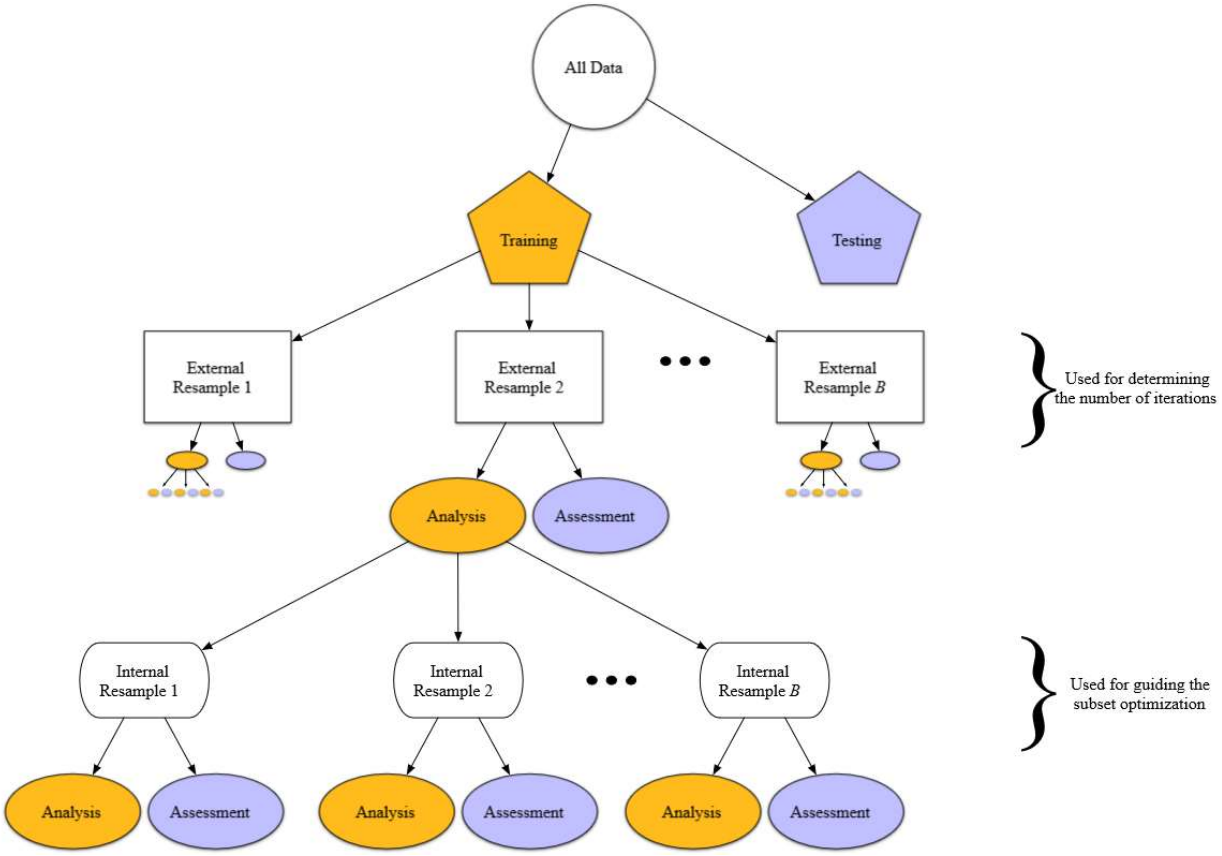
- *i* = iteration
- *c* = constant
- Once the probability is calculated, randomly choose a number between 0 and 1 from a uniform distribution and compare to probability.
- If random number is greater, the new feature set is rejected

More on Simulated Annealing

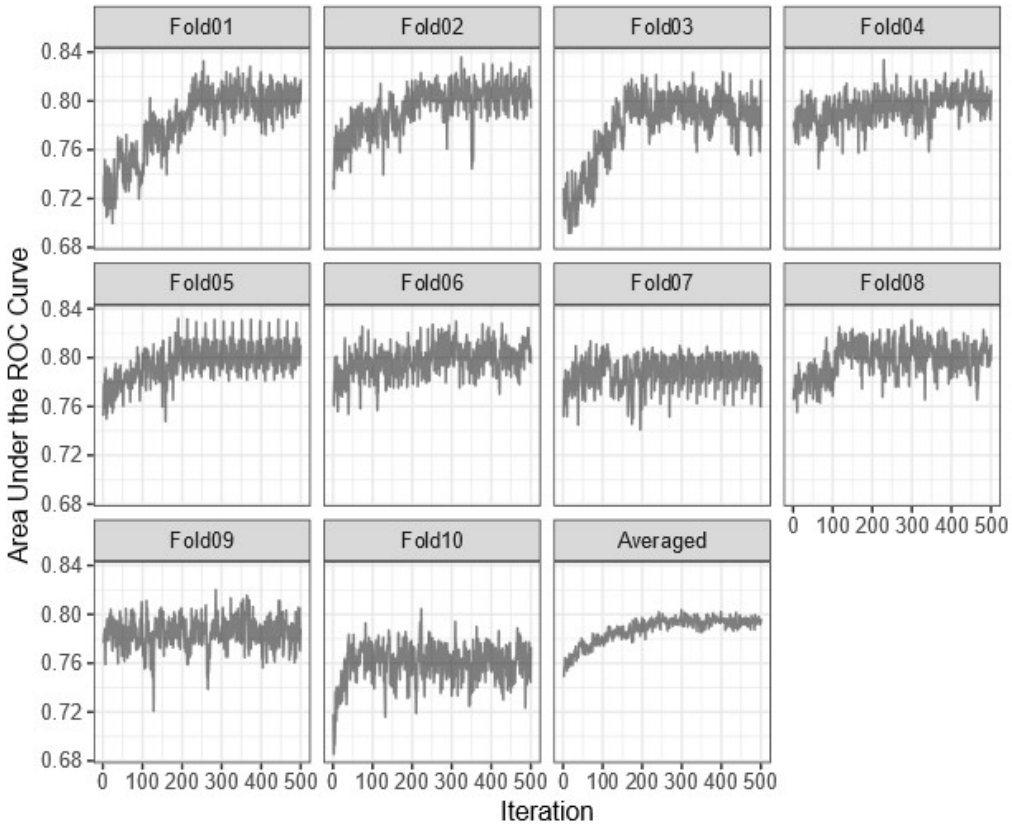
- If an optimal solution has not been found within *I* iterations, search resets to the last known optimal solution and proceeds again with *I* being the number of iterations since the restart

Iteration	Size	ROC	Probability	Random Uniform	Status
1	122	0.776	—	—	Improved
2	120	0.781	—	—	Improved
3	122	0.770	0.958	0.767	Accepted
4	120	0.804	—	—	Improved
5	120	0.793	0.931	0.291	Accepted
6	118	0.779	0.826	0.879	Discarded
7	122	0.779	0.799	0.659	Accepted
8	124	0.776	0.756	0.475	Accepted
9	124	0.798	0.929	0.879	Accepted
10	124	0.774	0.685	0.846	Discarded
11	126	0.788	0.800	0.512	Accepted
12	124	0.783	0.732	0.191	Accepted
13	124	0.790	0.787	0.060	Accepted
14	124	0.778	—	—	Restart
15	120	0.790	0.982	0.049	Accepted

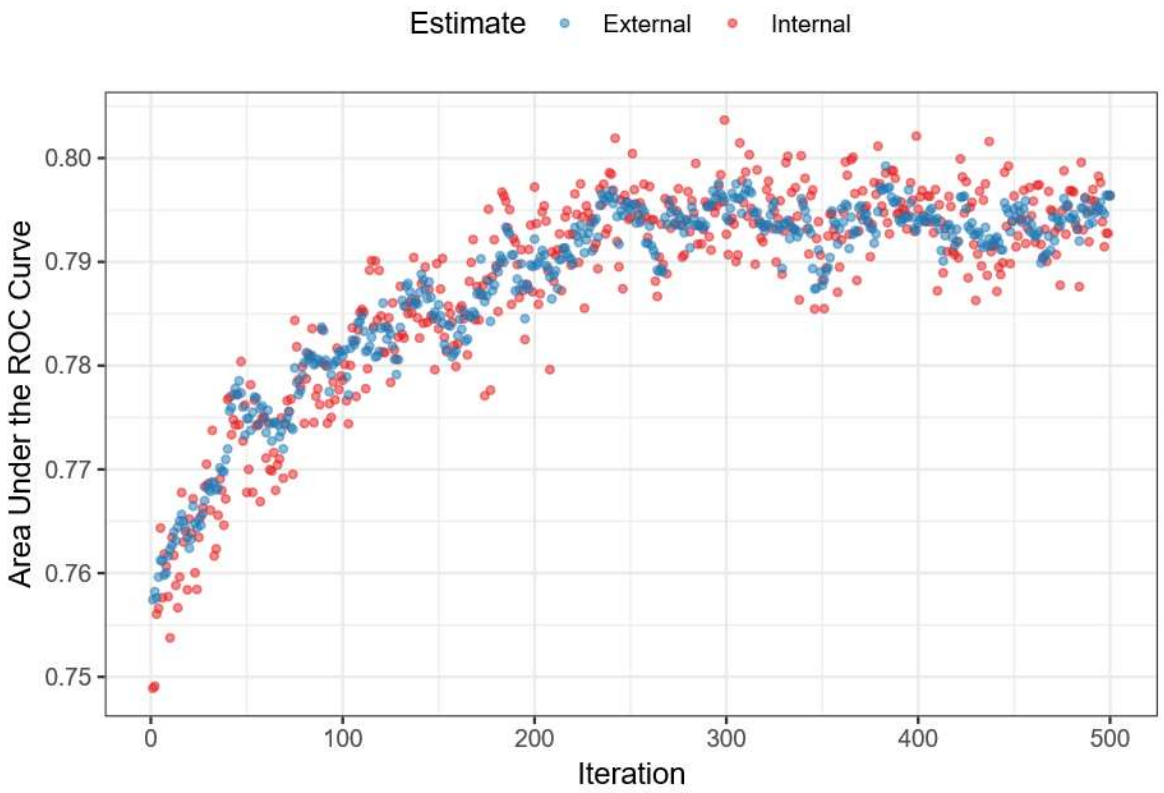
Resampling


All Data
Training
Testing
External Resample 1
External Resample 2
External Resample B
Used for determining the number of iterations
Analysis
Assessment
Internal Resample 1
Internal Resample 2
Internal Resample B
Used for guiding the subset optimization
Analysis
Assessment
Analysis
Assessment
Analysis
Assessment

Looking at the Folds



OkCupid


Estimate
External
Internal
Area Under the ROC Curve
0.80
0.79
0.78
0.77
0.76
0.75
0
100
200
300
400
500
Iteration

Final Check

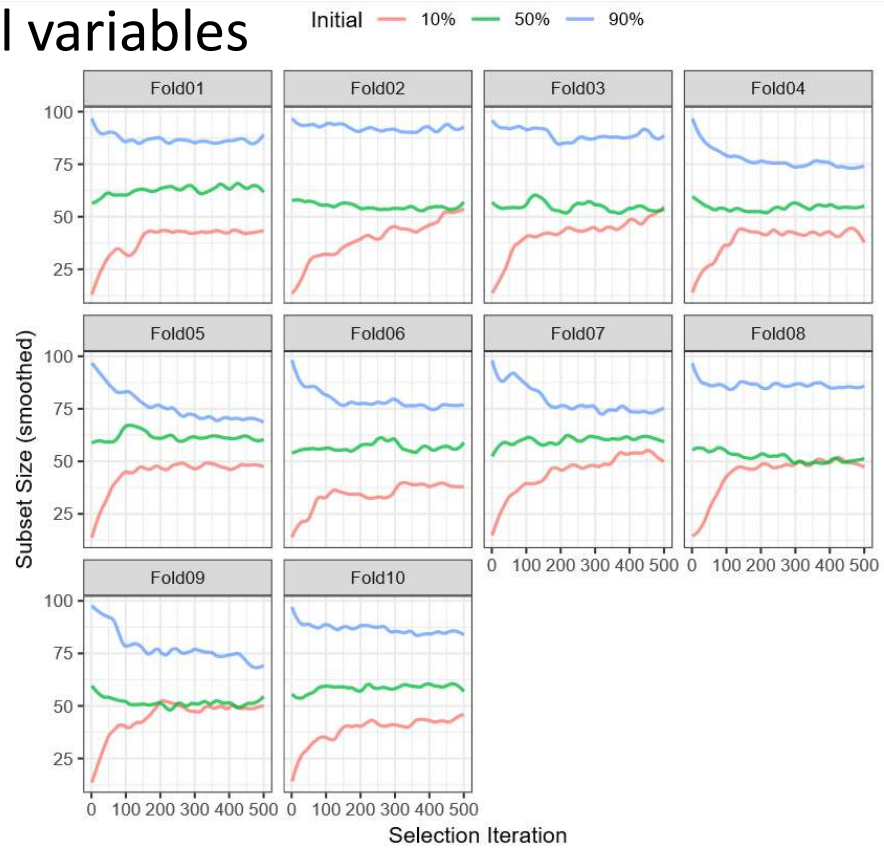
- A randomization approach was used to investigate if the 52 selected features contained good predictive information. For these data, 100 random subsets of size 66 were generated and used as input to naive Bayes models. The performance of the model from SA was better than 95% of the random subsets. This result indicates that SA selection process approach can be used to find good predictive subsets.

t-test

- “For a SA search lasting 500 iterations, there should be roughly 250 subsets with and without each predictor and each of these has an associated area under the ROC curve computed from the external holdout set.”
- Most significant variables from t-test: education level, software, keyword, income, the number of words, startup, solving, the number of characters/word, height, and the number of lower space characters

Final Notes on SA

- Could dummy code categorical variables



Genetic Algorithms

Example

ID										Performance	Probability (%)
1			C			F				0.54	6.4
2	A			D	E	F		H		0.55	6.5
3				D						0.51	6.0
4					E					0.53	6.2
5				D			G	H	I	0.75	8.8
6		B			E		G		I	0.64	7.5
7		B	C			F			I	0.65	7.7
8	A		C		E		G	H	I	0.95	11.2
9	A		C	D		F	G	H	I	0.81	9.6
10			C	D	E				I	0.79	9.3
11	A	B		D	E		G	H		0.85	10.0
12	A	B	C	D	E	F	G		I	0.91	10.7

Babies

ID								
6		B			E		G	I
12	A	B	C	D	E	F	G	I

The resulting children of these parents would be:

ID								
13		B			E	F	G	I
14	A	B	C	D	E		G	I

Mutation?

Natural Selection

- A logical approach to selecting parents would be to choose the top-ranking features subsets as parents. However, this greedy approach often leads to lingering in a locally optimal solution. To avoid a local optimum, the selection of parents should be a function of the fitness criteria. The most common approach is to select parents is to use a weighted random sample with a probability of selection as a function of predictive performance.

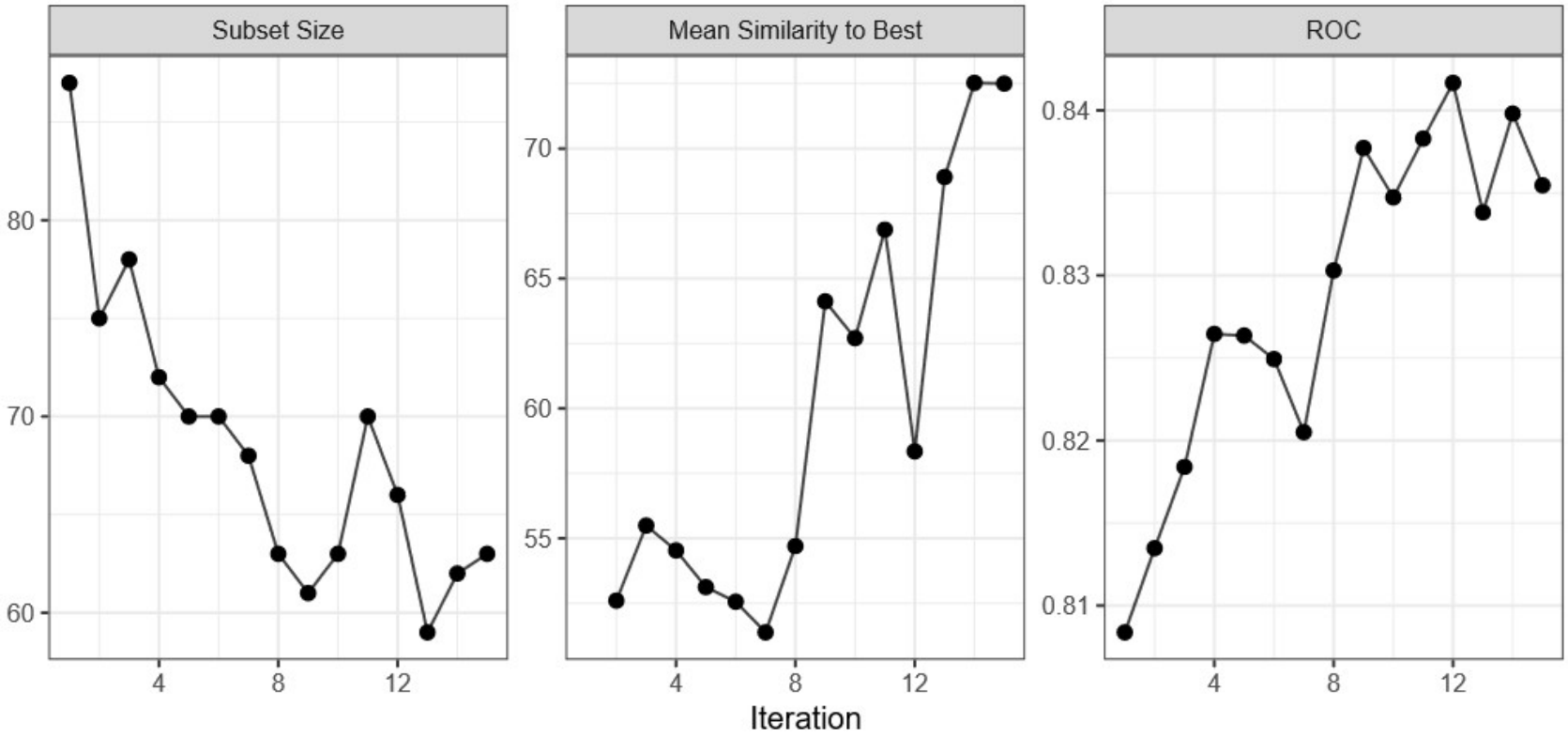
OkCupid

- Generation size: 50,
- Crossover probability: 80%,
- Mutation probability: 1%,
- Elitism: No,
- Number of generations: 14

Jaccard similarity

- This metric is the proportion of the number of predictors in common between two subsets divided by total number of unique predictors in both sets. For example, if there were five possible predictors, the Jaccard similarity between subsets ABC and ABE would be 2/4 or 50%. This value provides insight on how efficiently the generic algorithm is converging towards a common solution or towards potential overfitting.

Results


Subset Size
Mean Similarity to Best
ROC
80
70
60
70
65
60
55
0.84
0.83
0.82
0.81
4
8
12
Iteration

GA Effectiveness

- To gauge the effectiveness of the search, 100 random subsets of size 63 were chosen and a naive Bayes model was developed for each. The GA selected subset performed better than 100% of the randomly selected subsets. This indicates that the GA did find a useful subset for predicting the response.

Desirability Functions

- A simple desirability function would be a line that has zero desirability when the AUC is less than or equal to 0.50 and is 1.0 when the AUC is 1.0. Conversely, a desirability function for the number of predictors in the model would be 0.0 when all of the predictors are in the model would be 1.0 when only one predictor is in the model.
- Once all of the individual desirability functions are defined, the *overall desirability statistic* is created by taking the *geometric mean* of all of the individual functions.