The confection

m&m's
(185g)

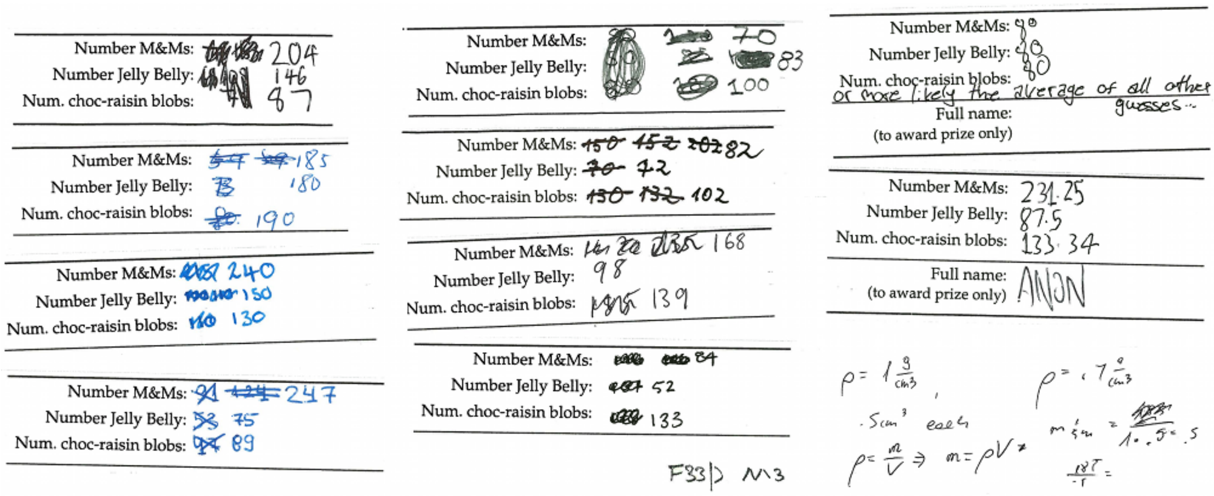
Jelly Belly
(100g)

Chocolate Raisins
(200g)

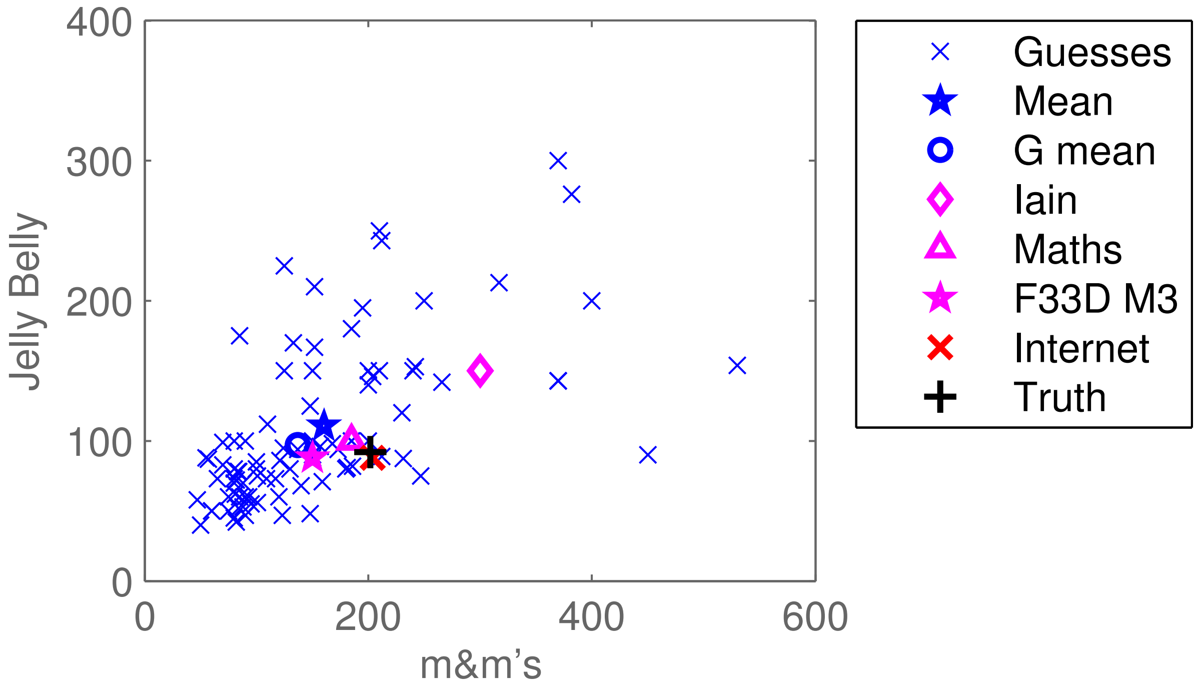
The importance of guessing

http://StreetFightingMath.com/

Stuff Inf2b students wrote

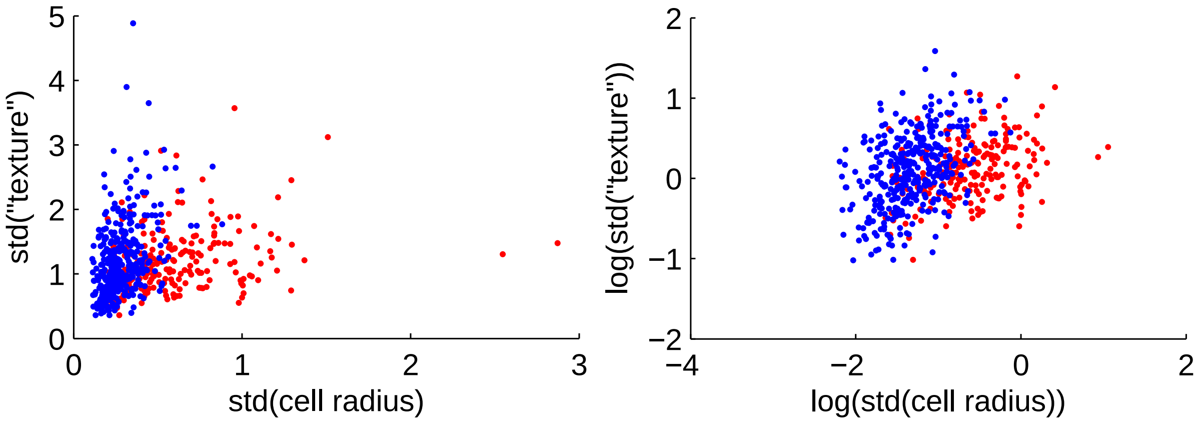


A 2D space



For 3D and more, check out the code on the website.

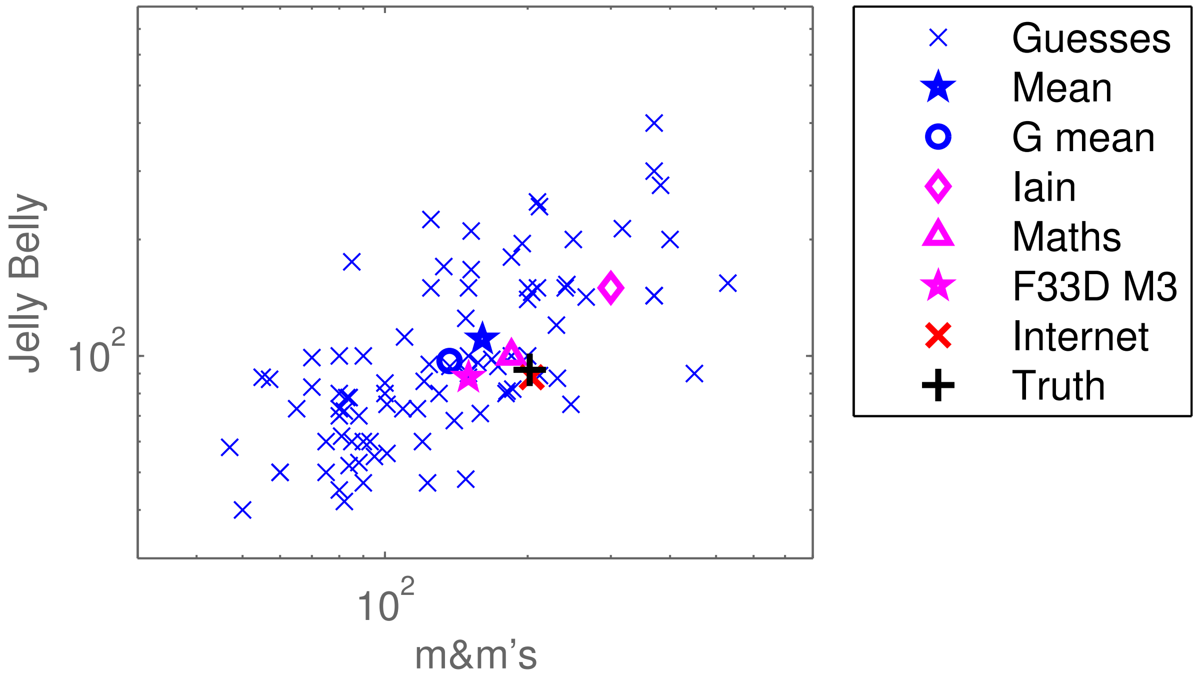
Often log-transform +ve data



Wisconsin breast cancer data

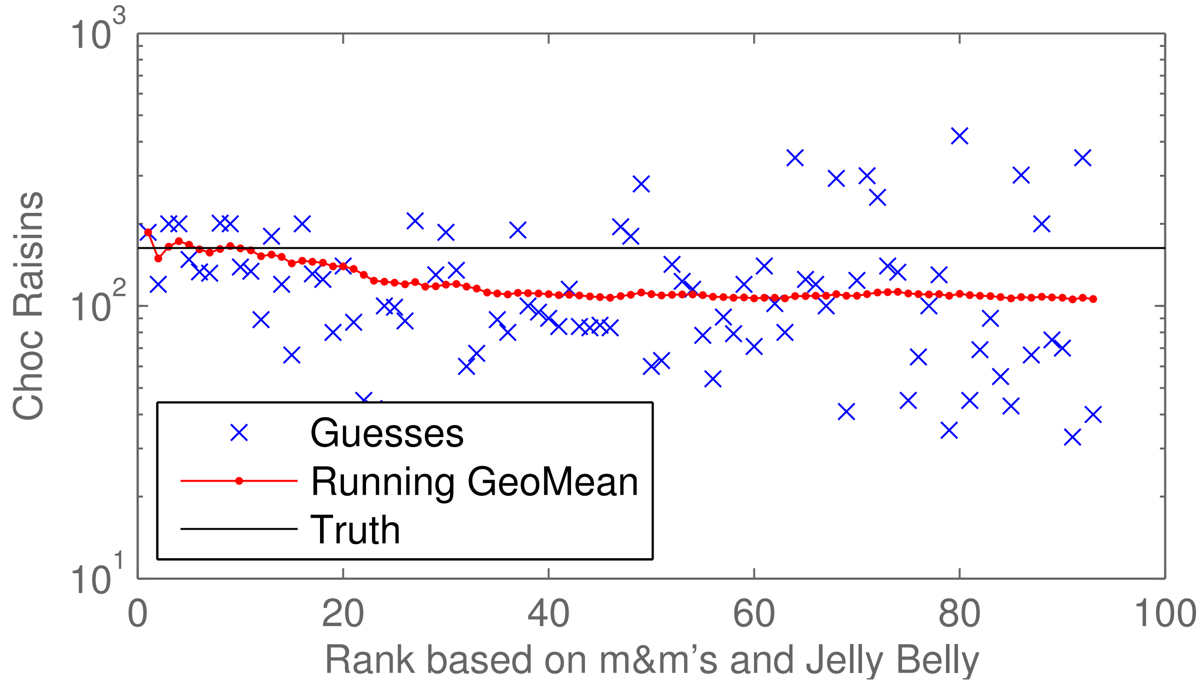
UCI ML repository

Count guesses on log-scale



Were some people just lucky?

Ranking by past performance

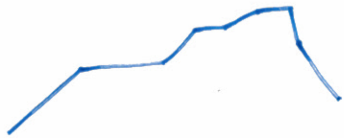


Ensemble of Models

Ensembles can ① reduce overfitting

and. can ② reduce underfitting

Build complicated $f^{\underline{n}}$



from simple pieces

Eg of ①

Bayesian model averaging:

$$p(y \mid \underline{x}, D) = \int p(y \mid \underline{x}, \underline{w})\, p(\underline{w} \mid D)\, d\underline{w}$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} p(y \mid \underline{x}, \underline{w}^{(s)}) \qquad \underline{w}^{(s)} \sim p(\underline{w} \mid D)$$

Ensemble of S predictors

What about making point prediction or guesses?

Kaggle give you a loss fn

$$L(\hat{y}; y) \underset{eg}{=} (\hat{y}-y)^2 \text{ or } |\hat{y}-y|$$

$\hat{y}$: guess, y: answer, or . . .

Then minimize expected loss:

$$\text{Minimize } \mathbb{E}_{p(y|\underline{x},D)}[L(\hat{y};y)]$$

$$= \int p(y|\underline{x},D)\, L(\hat{y};y)\, dy$$

For square error:

$$\Rightarrow \hat{y} = \text{mean of } p(y|\underline{x},D)$$

$$\text{or } \mathbb{E}_{p(y|\underline{x},D)}[y]$$

For absolute error:

$$\Rightarrow \hat{y} = \text{median of } p(y|\underline{x},D)$$

Another way of averaging predictions: Bagging

Bootstrap aggregation

You have a dataset of N examples

Train time

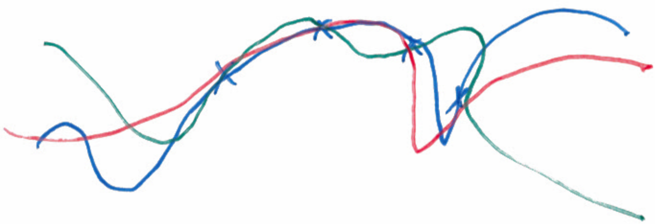
for $s = 1 \ldots S$

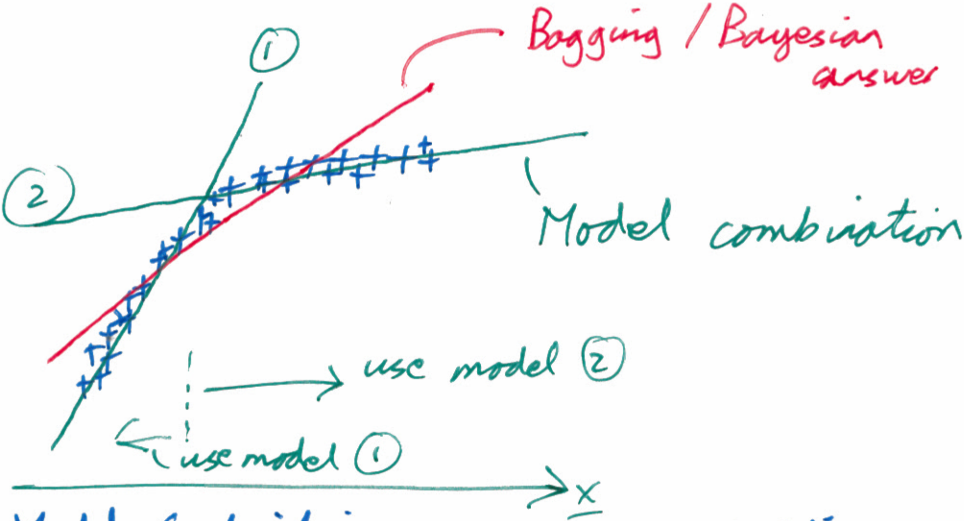
Create a new dataset of N examples by sampling with replacement from training set. (Bootstrap)

Fit your model → predictor s

Test time

Average predictions of S models





Model Combination

$$p(y|\underline{x}, \theta) = \sum_k \underbrace{p(y|\underline{x}, z{=}k, \theta)}_{\text{predictor/experts}} \underbrace{p(z{=}k|\underline{x}, \theta)}_{\text{classifier}}$$

params

$k \in \{1, 2, \ldots\}$

"Mixture of experts"

Fit θ, Max. Likelihood, + regularize.

Bayesian → Laplace / Var approx of $p(\theta|D)$

→ Or Bagging.

I showed you Bucilă et al. and Caruana et al. papers. See links in notes.