
NETFLIX
Netflix Prize
Home
Rules
Leaderboard
Register
Update
Submit
Download
NETFLIX
Browse
Recommendations
Friends
Queue
Buy DVDs
New Releases
Previews
Netflix Top 100
You really liked it...
Welcome!
The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.
Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.
You should also read the frequently-asked questions about the Prize. And check out how various teams are doing on the Leaderboard.
Good luck and thanks for helping!
FAQ | Forum | Netflix Home
© 1997-2006 Netflix, Inc. All rights reserved.

L16 (10) 2018

Is there any customer information in the dataset that should be kept private?

No, all customer identifying information has been removed; all that remains are ratings and dates.

This follows our privacy policy, which you can review here.

Even if, for example, you knew all your own ratings and their dates you probably couldn't identify them reliably in the data because only a small sample was included (less than one-tenth of our complete dataset) and that data was subject to perturbation.

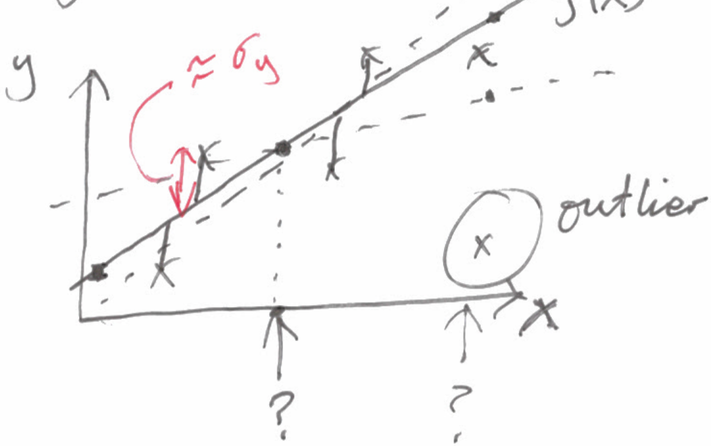
Of course, since you know all your own ratings that really isn't a privacy problem is it?

	Sound of music	Terminator	. . .	XXX
Customers				
~~Fred~~				
. . .				
86713 ~~Iain~~	5	4	2	2 5 0 0
. . .				
⋮				

IMDB

Iain	5	5	2	1 0 0 0

Bayesian Regression



f(x) is a single guess of output.

For classification we fitted

$$P(y \mid \underline{x}) \text{ by max. likelihood.}$$

For regression, write down a probabilistic model!

$$p(y \mid \underline{x}) = N(y; f(\underline{x}; \underline{w}), \sigma_y^2)$$

↑ noise variance

Simplest assumption: σ_y^2 known, same for every output.

Maximum Likelihood (not Bayesian)

Minimize negative log-likelihood

$$-\log p(\underline{y} \mid X, \underline{w}) = -\sum_n \log p(y^{(n)} \mid \underline{x}^{(n)}, \underline{w})$$

$\underline{y}$: $N \times 1$

$$= \sum_n \frac{1}{2\sigma_y^2}\left(y^{(n)} - f(\underline{x}^{(n)}; \underline{w})\right)^2 + \sum_{n=1}^{N} \frac{1}{2} \log 2\pi\sigma_y^2$$

$$= \frac{1}{2\sigma_y^2} \underbrace{\sum_n \left(y^{(n)} - f(\underline{x}^{(n)}; \underline{w})\right)^2}_{\text{Minimize}} + \frac{N}{2} \log 2\pi\sigma_y^2$$

$\Rightarrow$ ML for this model is least squares

Other costs

Variable noise: $\sigma_y^{(n)}$ for each data point.

$\Rightarrow$ Weight each example by $\frac{1}{(\sigma_y^{(n)})}$

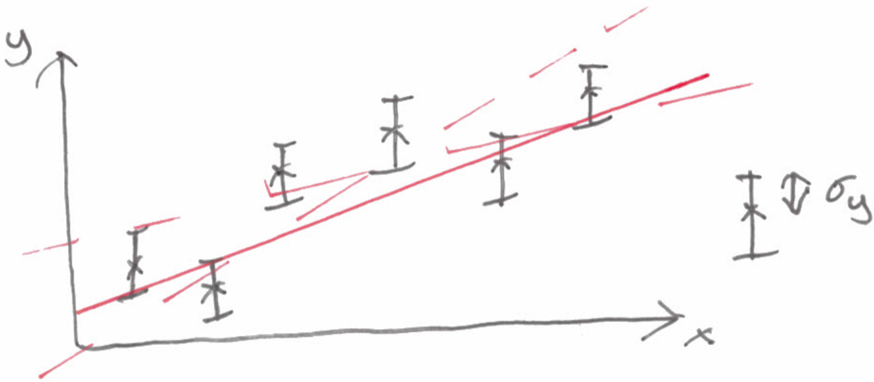
Robust model

As before: $p(y|\underline{x}, \underline{w})$ is a mixture between usual model (narrow Gaussian) and a "background model"

maybe $N(0, 10^4)$

Derive Likelihood → cost → gradients.

We are uncertain about a model given data



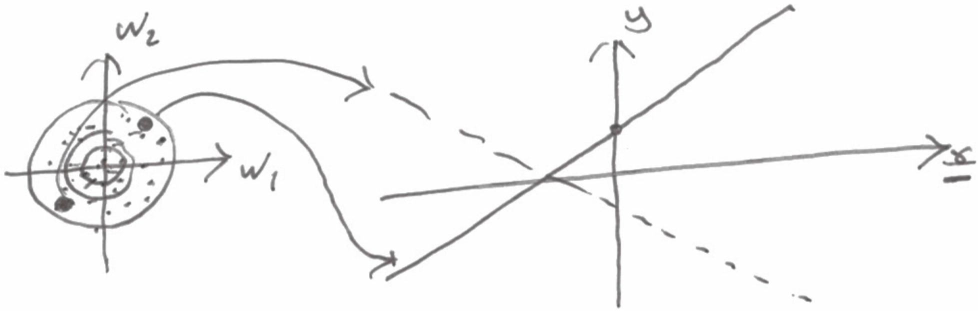
How do we automate?

Bayesian Probability Theory

Prior distribution

What model parameters are plausible?

$$p(\underline{w}) = \mathcal{N}(\underline{w}; \underline{0}, \sigma_w^2 \mathbb{I})$$

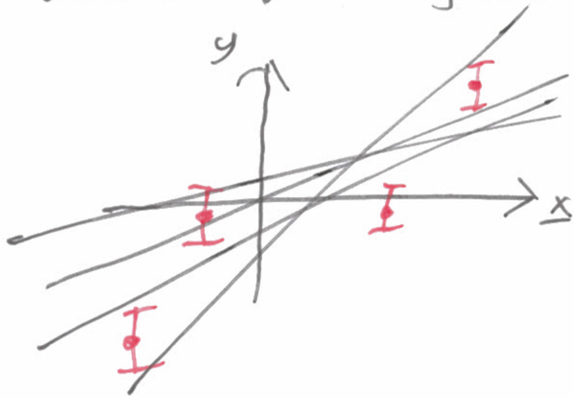


$$f = w_1 x + w_2$$

Bayes' Rule $\Downarrow$

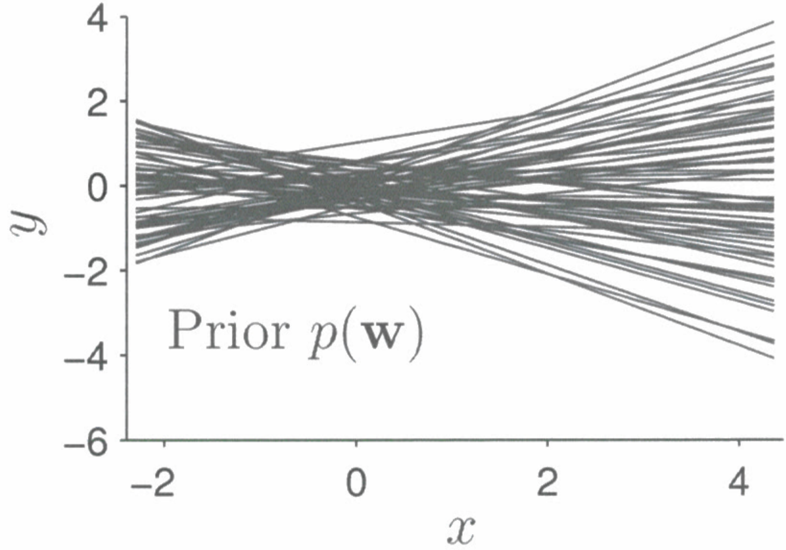
Updates these beliefs using data

$p(\underline{w} \mid Data)$



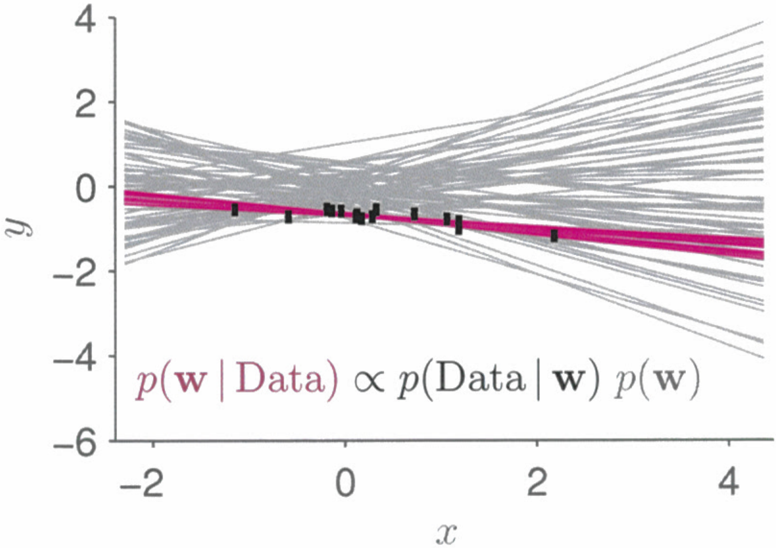
Linear regression

$y = w_1 x + w_2, \qquad p(\mathbf{w}) = \mathcal{N}(\mathbf{w};\, 0, 0.4^2 I)$

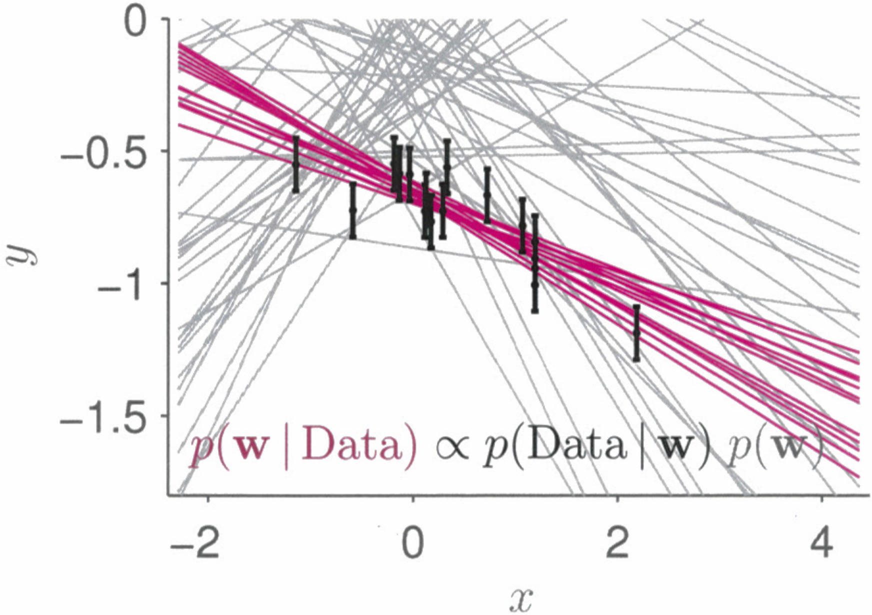


Linear regression

$$y^{(n)} = w_1 x^{(n)} + w_2 + \epsilon^{(n)}, \qquad \epsilon^{(n)} \sim \mathcal{N}(0, 0.1^2)$$



Linear regression (zoomed in)



Mid-Semester Survey

tinyurl.com/

mlpr2018mss