Machine Learning

4771

Instructor: Itsik Pe'er

About me

- Itsik Pe’er, Computational Geneticist
- Contact: CSB 505 (enter through MUDD 4^{th})
- Office hours: 5:35-6:35 Wed (ML) & most Mon
  - If you can’t get in: 212-9397135
  - In case of special issues or conflict w/ times: itsik@cs.columbia.edu

Staff

- Kristy Choi
- Eugene Ang
- Vidya Venkiteswaran
- Zhenrui Liao
- Antonio Moretti
- Alan Duan
- Rong Zhou

>Daily office hours, listed on a file on courseworks/Admin
Online on Piazza
ml4771tas@lists.cs.columbia.edu
Individual emails listed on a file on courseworks/Admin

Why this class?

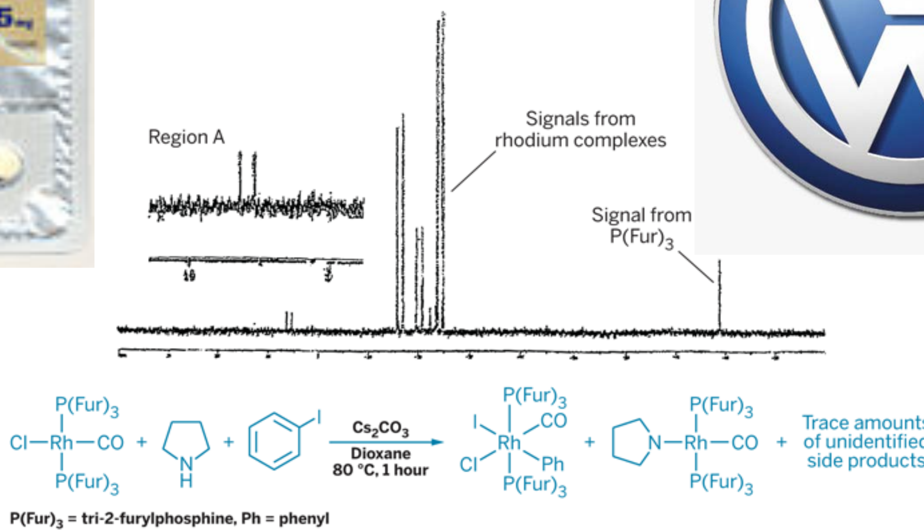
- Exciting times for ML
- Approach: fishing rods (understand methods) not just fish (apply tools blindly)
- Collective wisdom

Class #1

- Introductions & administration
  - Syllabus, policies, texts, courseworks
- Machine Learning: what, why and what for
  - Historical Perspective
  - Machine Learning Tasks, Tools & Approaches
  - Example

Academic Honesty

Reflects our social responsibility as engineers and data scientists



Academic Honesty

Reflects our social responsibility as engineers

- May be different than your home department
- You can discuss with a disclosed partner
- You must write/code up your own homework
- Use public libraries legally, as directed
- Don’t copy code or work by others
- No collaboration on quizzes, midterm, & final
- Assignments will be checked for plagiarism
- Class policy is t0 refer all cases to the Dean

Waiting List Policy

- In-class: Now at capacity
- Based on need & background
  - Send requests to ml4771tas@lists.cs.columbia.edu
- Hybrid section: ~all eligible admitted
- See enrollment FAQ http://bit.ly/2jx1VxY

What you need to know coming in

- Probability (statistics)
  - Definitions (probability space, events, conditional p, random variables), distributions (discrete & continuous, 1- & multi-D, Bernoulli, uniform, binomial, geometric, exponential, Poisson, normal), moments (expectation, variance, standard deviation, correlation) theorems (large numbers, central limit)
  - Review on Monday + HW0
- Lin. Algebra: matrices, eigenvalues
- Calc: multi-D differential & intergral

Course Details & Requirements

- Reference Text: Pattern Recognition & Machine Learning
  by C. Bishop (Spring 2006 Edition)
- Later in class: Probabilistic to Graphical Models
  by D. Koller & N. Friedman (1^{st} Edition)

- Homework: Every 7-14 days; submit what you have on time.
- Grade: HW (25%), midterm (25%), 2xquiz (20%)& final exam
- Appeals: within 2 weeks

- Software requirements: Python
- Class Google Cloud for resource-intensive assignments later

Courseworks Page

Slides will be available on courseworks

Link to videos

Check courseworks regularly for readings, homework deadlines, announcements, etc.

Submission: on courseoworks

General questions: Piazza

Schedule

- Feb 19: Quiz
- March 13, 15: Break
- March 22-24: Take-home midterm
- April 15: Quiz (incremental)
- May 8: Final

- See calendar on courseworks

Syllabus

- Week 1: Intro to ML
- Week 2: Review probability, regularized regression
- Week 3: Parameter estimation, multi-D Gaussians
- Week 4: Linear classification
- Week 5: SVMs
- Week 6: Kernels, decision trees
- Week 7: Nonlinear networks, back propagation
- Week 8: Nearest neighbors, dim. reduction
- Week 9: Review, midterm
- Week 10: Clustering, Gaussian mixtures
- Week 11: HMMs
- Week 12: Graphical models
- Week 13: Clique-tree Bayesian networks & causality
- Week 14: Cyclical dependencies, Markov Random Fields

Credit for much of the material: Jebara, Hsu

Machine Learning: What/Why

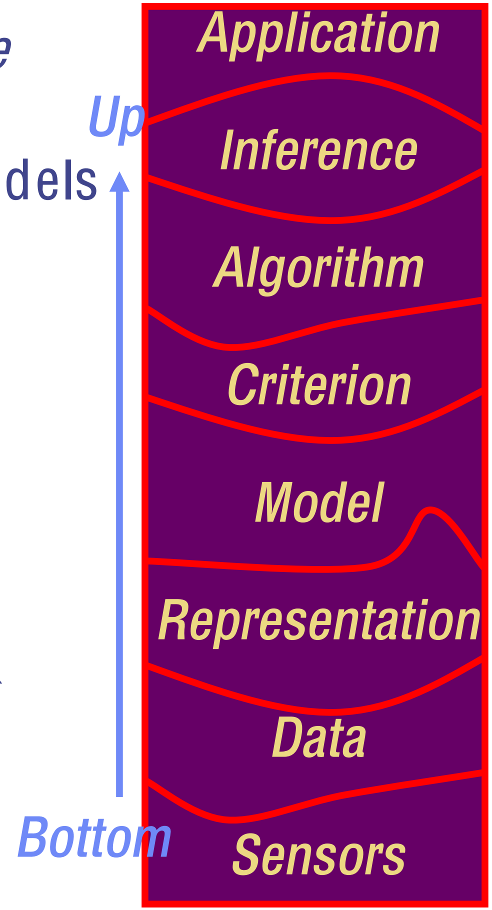
Algorithms that improve upon experience

Statistical Data-Driven Computational Models

Real domains (vision, speech, behavior):
- no $E=MC^2$
- noisy, complex, nonlinear
- have many variables
- non-deterministic
- incomplete, approximate models

Need: statistical models driven by data & sensors, a.k.a Machine Learning

Bottom-Up: use data to form a model



Intelligence = Learning = Prediction

Historical Perspective (Bio/AI)

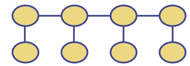
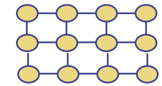
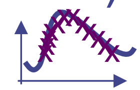
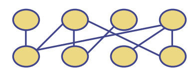
- 1917: Karel Capek (Robot)
- 1943: McCullogh & Pitts (Bio, Neuron)
- 1947: Norbert Weiner (Cybernetics, Multi-Disciplinary)
- 1949: Claude Shannon (Information Theory)
- 1950: Minsky, Newell, Simon, McCarthy (Symbolic AI, Logic)
- 1957: Rosenblatt (Perceptron)
- 1959: Arthur Samuel
  Coined Machine Learning
  Learning Checkers

- 1969: Minsky & Papert (Perceptron Linearity, no XOR)
- 1974: Werbos (BackProp, Nonlinearity)
- 1986: Rumelhart & McLelland (MLP, Verb-Conjugation)
- 1980’s: NeuralNets, Genetic Algos, Fuzzy Logic, Black Boxes

Historical Perspective (Stats)

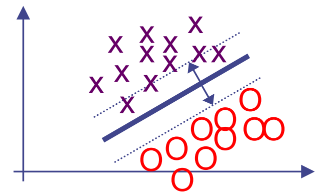
- 1763: Bayes (Prior, Likelihood, Posterior)
- 1920's: Fisher (Maximum Likelihood)
- 1937: Pitman (Exponential Family)
- 1969: Jaynes (Maximum Entropy)
- 1970: Baum (Hidden Markov Models)
- 1978: Dempster (Expectation Maximization)
- 1980's: Vapnik (VC-Dimension)
- 1990's: Lauritzen, Pearl (Graphical Models)
- 2000's: Bayesian Networks, Graphical Models, Kernels, Support Vector Machines, Learning Theory, Boosting, Active, Semisupervised, MultiTask, Sparsity, Convex Programming
- 2010's: Nonparametric Bayes, Spectral Methods, Deep Belief Networks, Structured Prediction, Conditional Random Fields

Current Applications

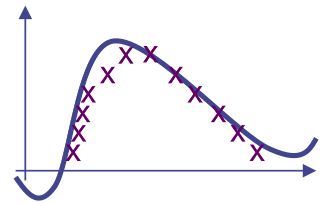
- Speech Recognition (HMMs, ICA)

- Computer Vision (face rec, digits, MRFs, super-res)

- Time Series Prediction (weather, finance)

- Genomics (micro-arrays, SVMs, splice-sites)
- NLP and Parsing (HMMs, CRFs, Google)
- Text and InfoRetrieval (docs, google, spam, TSVMs)
- Medical (QMR-DT, informatics)

- Behavior/Games (reinforcement, recommendations, SVD)
- Robotics (self-driving, workforce)

Machine Learning Tasks

Classification y=sign(f(x))

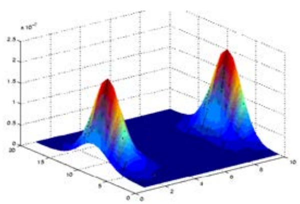


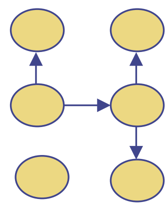
Regression y=f(x)



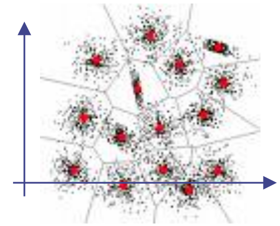
Supervised

Modeling p(x)

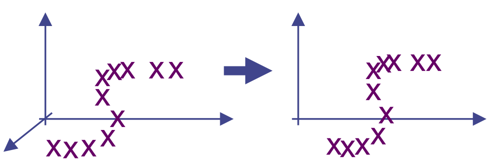




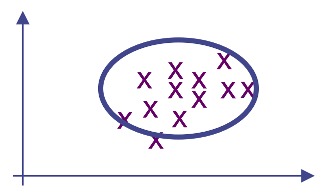
Clustering



Feature Selection



Detection p(x)<t



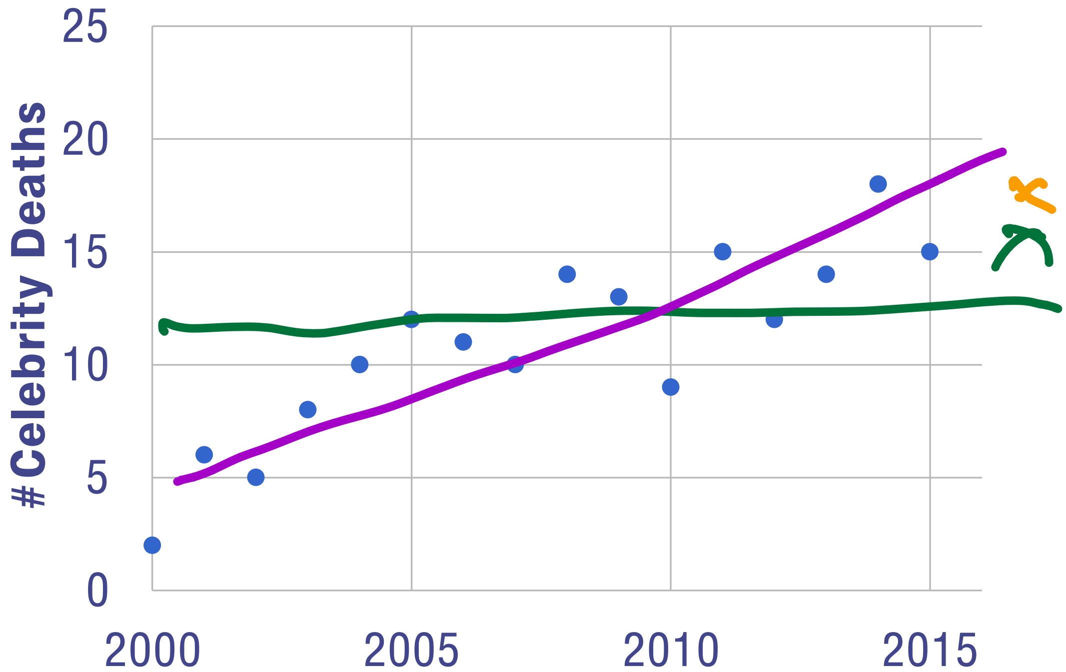
Unsupervised

Example: Celeb-lethality of 2016

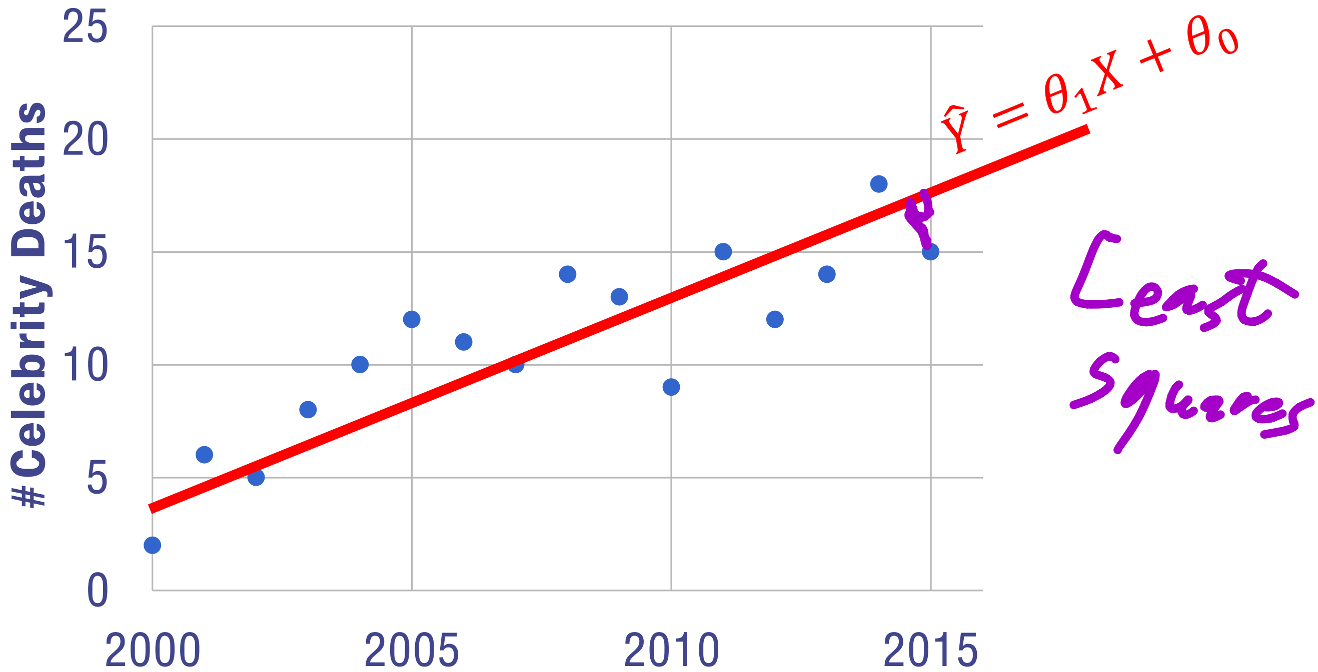
Did more celebrities die than what you would have predicted?

Itsik Pe'er, Columbia University

How many deaths are predicted?



How many deaths are predicted?

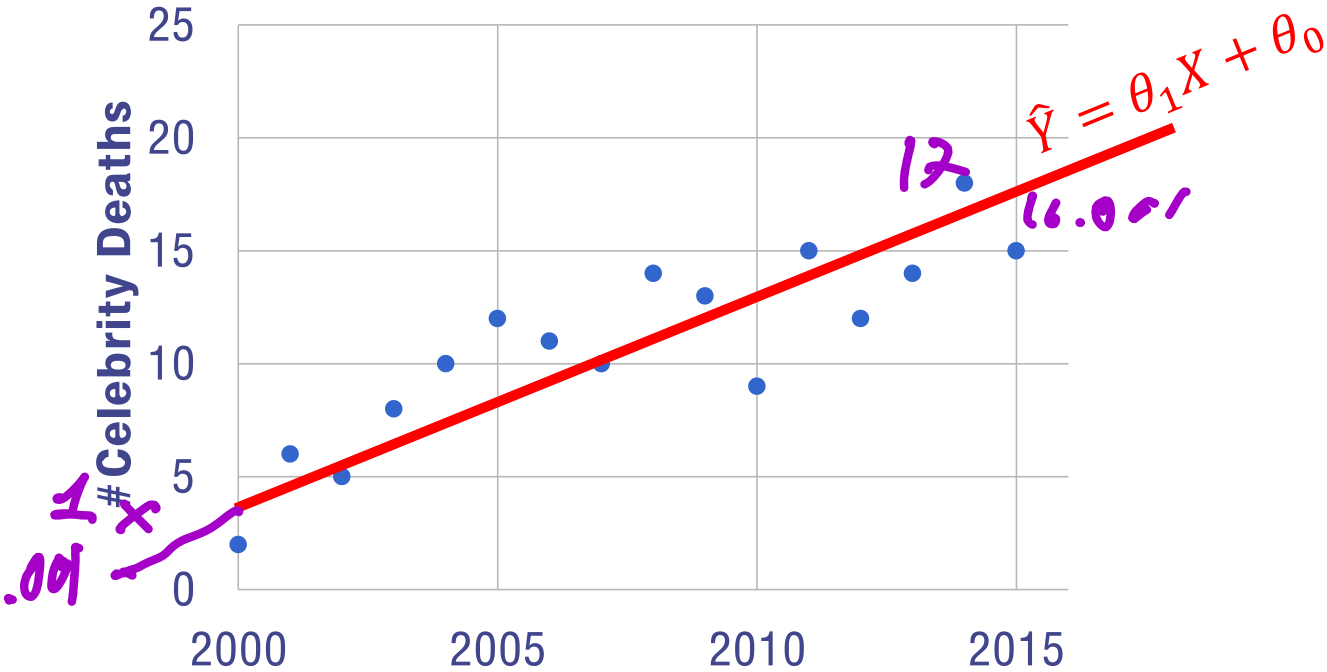


Find θ_1, θ_0 that best fit the observed Y

medium.com

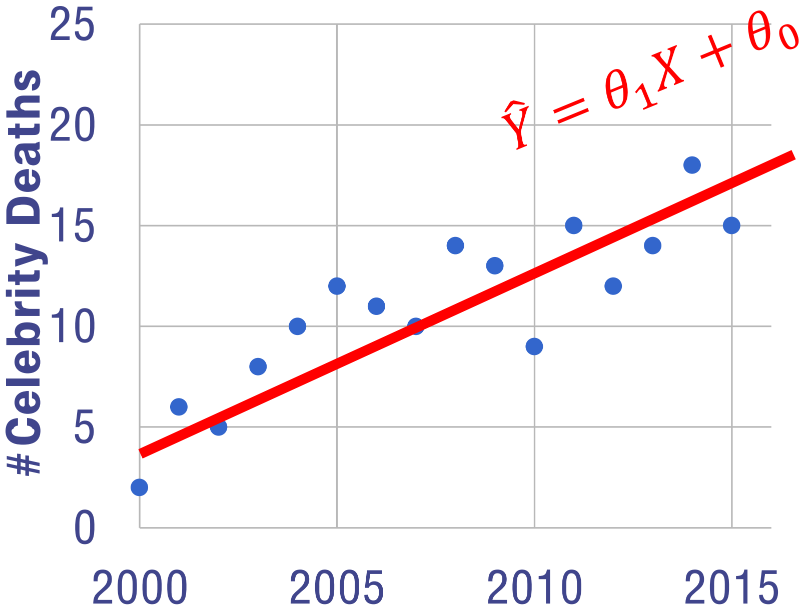
Questions

- Supervised or unsupervised?
- What does best fit mean?



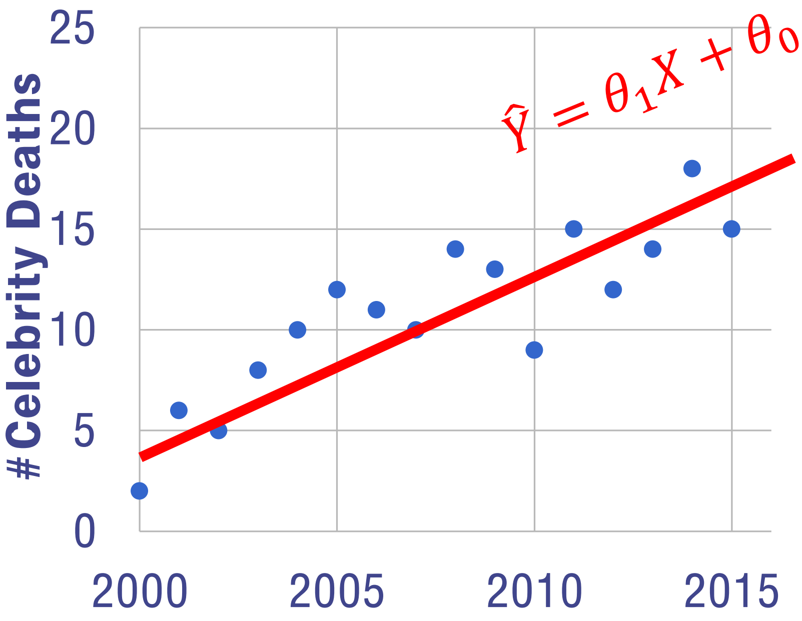
Probabilistic best-fit

- Best-fit =best at modeling data as plausible



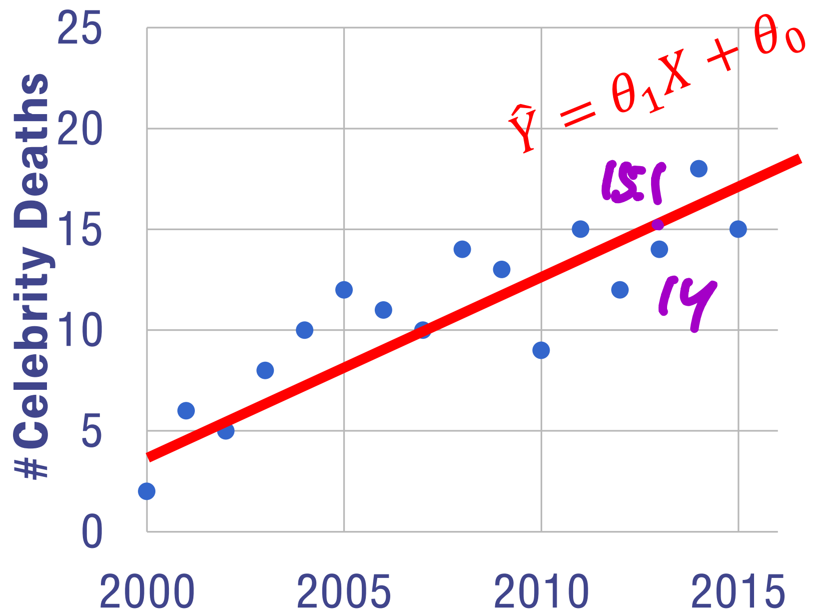
Probabilistic best-fit

- Best-fit = best at modeling data as plausible
- Likelihood of model: Prob(data|model)
- Find Max Likelihood



Probabilistic best-fit

- Best-fit = best at modeling data as plausible
- Likelihood of model: Prob(data|model)
- Find max likelihood
- Find θ_1, θ_0
  s.t. $\mathrm{Prob}(Y|\hat{Y})$
  is maximized
- What is $\mathrm{Prob}(Y|\hat{Y})$?



Digression/Review: Poisson

- Events at rate per $\lambda = \hat{Y}$ year

$$\text{Binomial}\left(n=12, p=\frac{\lambda}{12}\right)$$

$$\text{Binomial}\left(n=365, p=\frac{\lambda}{365}\right)$$

$$\lim_{n\to\infty} \text{Binomial}\left(n, \frac{\lambda}{\lambda}\right)$$

Digression/Review: Poisson

- Events at rate per $\lambda = \hat{Y}$ year

- $\text{Poisson}(\lambda) = \lim_{n\to\infty} \text{Binomial}(n, \frac{\lambda}{n})$
- $X \sim \text{Poisson}(\lambda)$:
  $\text{Prob}(X = k) = \lim_{n\to\infty} \cancel{\binom{n}{k}}^{\frac{n^k}{k!}} \left(\frac{\lambda}{\cancel{n}}\right)^k \cancel{\left(1-\frac{\lambda}{n}\right)}^{\cancel{n-k}} \quad e^{-\frac{\lambda}{\cancel{n}}(n \cancel{-k})}$

Digression/Review: Poisson

- Events at rate per $\lambda = \hat{Y}$ year

- $\text{Poisson}(\lambda) = \lim_{n\to\infty} \text{Binomial}(n, \frac{\lambda}{n})$
- X~Poisson(λ):

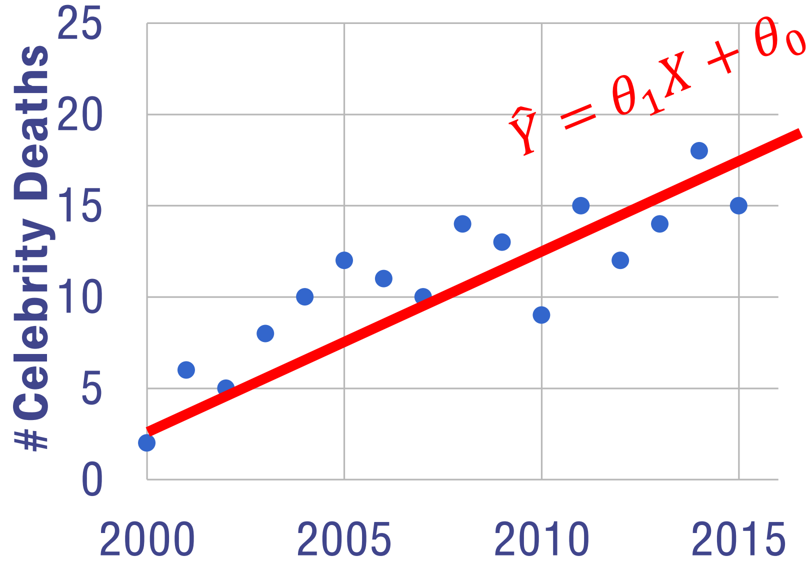
  $\text{Prob}(X = k) = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$
- $\text{Poisson}(k{:}\,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$

Probabilistic best-fit

- Maximize $L(\theta_1, \theta_0) = Prob(Y|\hat{Y})$

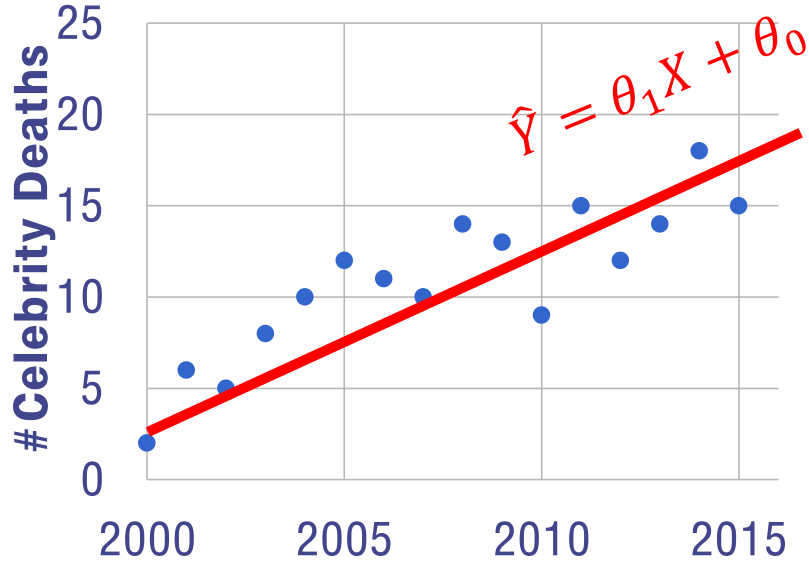
$$L(\theta_1, \theta_0) = \Pi_i Prob(y_i|\theta_1 x_i + \theta_0)$$

$$= \Pi_i \frac{(\theta_1 x_i + \theta_0)^{y_i} e^{-(\theta_1 x_i + \theta_0)}}{y_i!}$$



Probabilistic best-fit

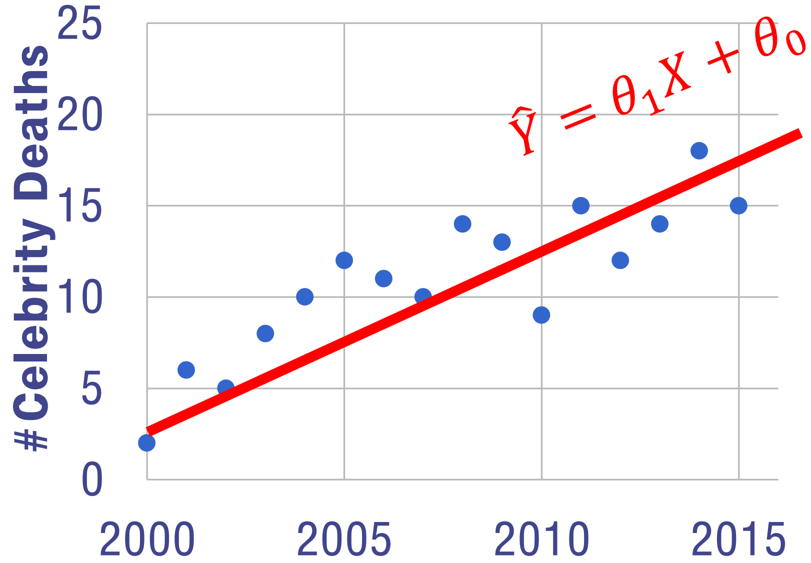
- Maximize $L(\theta_1, \theta_0) = \text{Prob}\left(Y \middle| \hat{Y}\right)$



Probabilistic best-fit

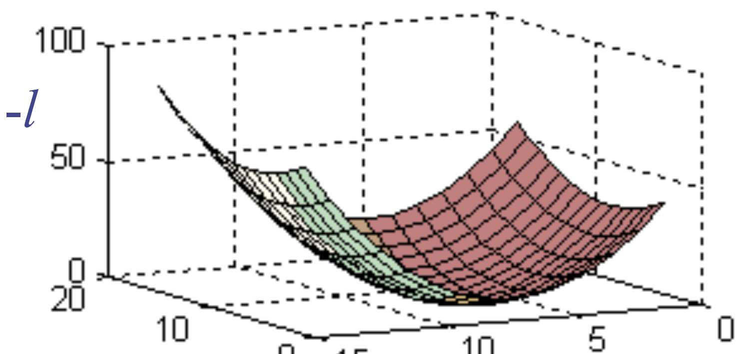
- Maximize $L(\theta_1, \theta_0) = \mathrm{Prob}\left(Y \middle| \hat{Y}\right)$
- $L(\theta_1, \theta_0) = \prod_i \mathrm{Prob}(y_i | \theta_1 x_i + \theta_0)$

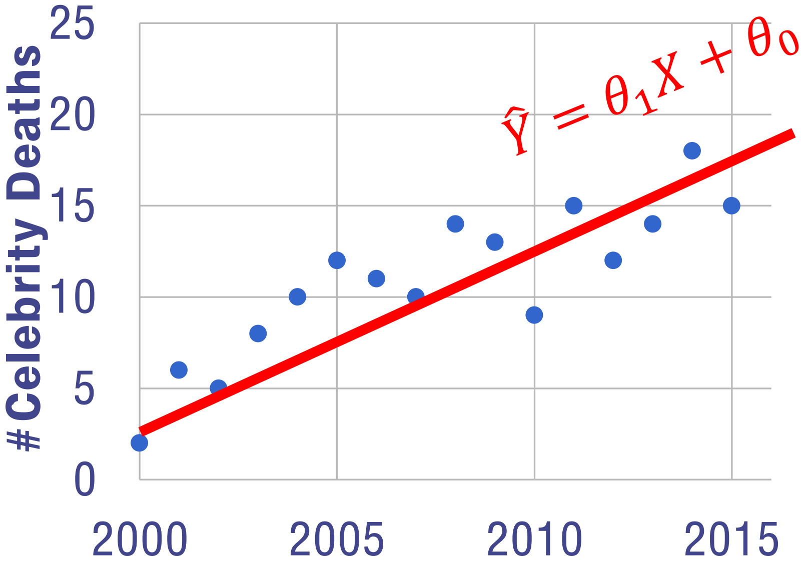
$$= \prod_i \frac{(\theta_1 x_i + \theta_0)^{y_i} e^{-(\theta_1 x_i + \theta_0)}}{y_i!}$$



Maximizing Likelihood

- $L(\theta_1, \theta_0) = \prod_i \frac{(\theta_1 x_i + \theta_0)^{y_i} e^{-(\theta_1 x_i + \theta_0)}}{y_i!}$
- $l(\theta_1, \theta_0) = \log L(\theta_1, \theta_0) =$
  $C + \sum_i [y_i \log(\theta_1 x_i + \theta_0) - (\theta_1 x_i + \theta_0)]$





Summary

- Welcome to Intro to Machine Learning
- Regression:
  - Fitting a probabilistic model to the data
  - Max likelihood