COMP 543: Tools & Models for Data Science

Intro to Modeling 2

Chris Jermaine & Risa Myers

Rice University

Models Are Parameterized

- Normal: μ, σ
  - μ is the center of the distribution
  - σ determines the width

$$f_{\text{Normal}}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Exponential: λ

$$f_{Exp}(x|\lambda) = \lambda e^{-\lambda x}$$

- Key question: how to choose parameters?

Models Are Parameterized

- Normal: μ, σ
  - μ is the center of the distribution
  - σ determines the width

$$f_{\text{Normal}}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Exponential: λ

$$f_{Exp}(x|\lambda) = \lambda e^{-\lambda x}$$

- Key question: how to choose parameters?
  - Typically chosen to “fit” the model to example data
  - To make the model a good explanation for the data
  - Also called “learning” in ML

Approaches to Learning a Model

- Are many, including:
  - Optimization based (ex: least squares)
  - Probabilistic: MLE (Maximum Likelihood Estimation)
  - Probabilistic: Bayesian
  - Deep Learning

Choosing an Approch to Learning a Model

- Nature of the problem
- Amount of data available
- Tools available
- Requirements of the model
  - Interpretability
  - Availability
  - Familiarity
  - Experimentation
  - Accuracy

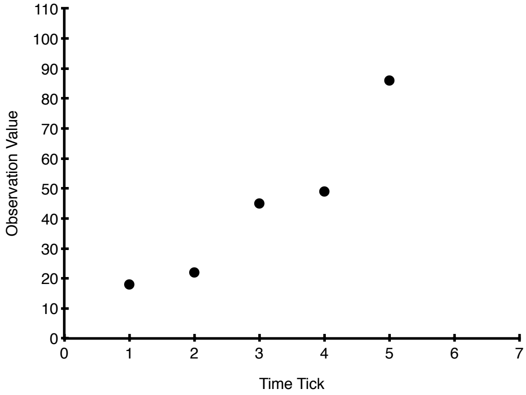
Optimization-Based

- Goal is to reduce some error metric on example/training data
  - Error tells us how well the model fits the TRAINING data
- No direct probabilistic motivation
- Common approach

Classic Example: Least Squares Regression

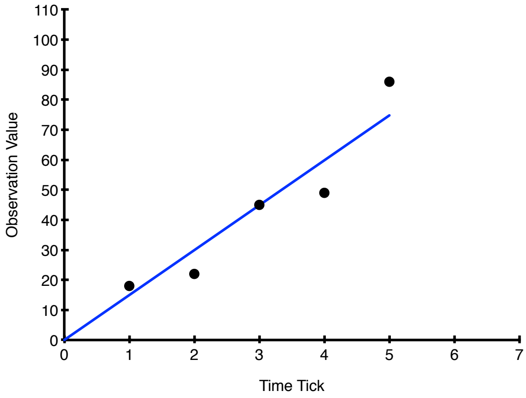
Example

- I observe $\langle 18, 22, 45, 49, 86 \rangle$
- At time ticks $\langle 1, 2, 3, 4, 5 \rangle$
- ? How can I predict the next item?



Example: Least Squares Regression

- I observe $\langle 18, 22, 45, 49, 86 \rangle$
- At time ticks $\langle 1, 2, 3, 4, 5 \rangle$
- How can I predict the next item?
- ith observation is x_i, tick is t_i
  - Might fit a line to the data
  - So, $x_i \approx f(t_i) = m \times t_i$
  - Where m is the slope of the line
  - Observation at time tick i is a function of t
  - t_i is the ith time tick
  - Here $t_i = i$ - our data is evenly spaced



Overlap of Approaches

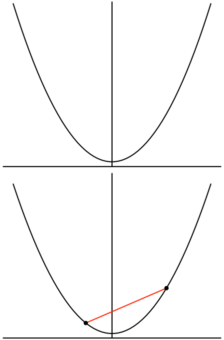
- Approaches may overlap
- Compare Least Squares to the PDF for the Normal distribution
  - Includes a term with $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - $(x-\mu)^2$ Looks a lot like Least Squares

Computing Least-Squares Fit

- Loss function is the sum of the squares of the residuals
- Defn: A residual is the difference between the observed value and the model computed value
- This loss function is “convex”

What is a Convex Function?

- Intuition
  - Continuous function
  - The line connecting any two points is on or above the function
  - The function isn't "wavy"
  - Strictly convex $\rightarrow$ 2nd derivative is always positive
- Examples
  - Quadratic function x^2
  - Exponential function e^x
- Benefits
  - Strictly convex functions have at most one minimum
  - Differentiable



Computing Least-Squares Fit

- Choose unique m where $l'(m) = 0$
- That is, where the derivative of the loss function = 0

Computing Least-Squares Fit

$$
\begin{aligned}
l(m) &= \sum_i (f(t_i) - x_i)^2 \\
&= \sum_i (m \times t_i - x_i)^2 \\
l'(m) &= \sum_i 2t_i (m \times t_i - x_i) \\
&= \sum_i 2mt_i^2 - 2t_i x_i \\
&= 2m(1+4+9+16+25) - 2(18+44+135+196+430) \\
&= 110m - 1646
\end{aligned}
$$

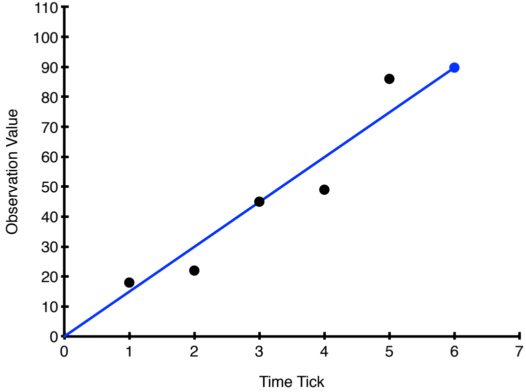
- Define loss function
- Sub in defn of m
- Take the first derivative (chain rule)
- Simplify
- Plug in values
- Solve for m

- So loss minimized at $m = 14.96$
- Recall $f(t_i) = m \times t_i$

? Value at time tick 6?

Next Value?

- Recall $f(t_i) = m \times t_i$
- Value at time tick 6?
- $f(6) = 14.96 \times 6 = 89.8$



Advantages and Disadvantages of Least Squares

?

Advantages and Disadvantages of Least Squares

- Advantages
  - Penalizes values as our predictions move further and further away from the observations
  - Mathematical convenience
- Disadvantages
  - Very sensitive to outliers
  - May require pre-processing / “cleaning” before it can be used

Other Loss Functions

- View the list of prediction errors $(f(t_i) - x_i)$ as a vector
- Can have many loss functions, corresponding to norms
- Given a vector of errors $\langle \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \rangle$, l_p norm defined as:

$$||l||_p = \left(\sum_{i=1}^{n} |\varepsilon_i|^p \right)^{1/p}$$

- Norm maps a vector to a non-negative scalar
- Reflect different "distance" measures

Other Loss Functions

- Common loss functions correspond to various norms:
  - l_1 corresponds to mean absolute error
    - Used in LASSO
  - l_2 to mean squared error/least squares
  - l_∞ corresponds to minimax, the max absolute value of the dimensions

Maximum Likelihood

- Often we have a proper stochastic model
- Ex: observed $\{18, 22, 45, 49, 86\}$
- Model is Exponential, unknown λ
- How to estimate?
  - Most commonly: perform MLE
  - Ignoring time ticks (for now)
- You choose the parameters to maximize the likelihood of getting the observed data

Likelihood

- First, need the notion of a “likelihood function”
- Best illustrated with an example
  - In our case, $f(x_i|\lambda) = \lambda e^{-\lambda x_i}$
  - Recall that $f(x_i|\lambda)$ is the probability density of the ith point
  - So $f(x_1, x_2, ..., x_n|\lambda) = \prod_i \lambda e^{-\lambda x_i}$
  - Assume iid, so the density is a product.
  - This is a PDF

Likelihood

- A “likelihood function” simply turns the parametrization around
  - Instead of $f(x_i|\lambda)$, we have $f(\lambda|x_i)$
  - So $L(\lambda|x_1, x_2, ..., x_n) = \prod_i \lambda e^{-\lambda x_i}$
  - Now L measures the goodness of the parameter λ
  - And NOT how likely $x_1, x_2, ..., x_n$ are given the model

Maximum Likelihood Estimation - MLE

- Given $L(\Theta^1|D)$ (Θ is set of model params, D is data)...
  - The MLE $\hat{\Theta}$ for Θ is defined as the value such that

$$\forall \hat{\Theta}', \; L(\hat{\Theta}'|D) \leq L(\hat{\Theta}|D)$$

  - In other words: $\hat{\Theta}$ is the best possible set of parameters, given the available data
  - Note: closely related to least squares, for normally distributed data

? Why do we like it?

[1] Θ is commonly used to denote a set of parameters

MLE

- Given $L(\Theta|D)$ (Θ is set of model params, D is data)...
  - The MLE $\hat{\Theta}$ for Θ is defined as the value such that

$$\forall \hat{\Theta}', \; L(\hat{\Theta}'|D) \leq L(\hat{\Theta}|D)$$

  - Note: closely related to least squares!!
- Why do we like it?
  - Under many conditions, it is the Minimum-variance unbiased estimator ("MVUE")
  - Under many conditions, error is asymptotically normal
  - This allows us to talk about confidence bounds

Minor Digression: Sources of Error

- Two most concerning ones
- Bias
  - “Expected error”
  - How far off your are on Expectation
  - e.g. Guess a number, many times
  - If my average error is +2, this is an example of bias
- Variance
  - “Spread”
  - How far from correct your mean guess is

MLE Example

- Want to find the decay parameter in an exponential distribution
- Might start with an unbiased estimator with low variance
- Classic estimators are often unbiased, due to asymptotic properties
- Are these the best to use?
  - Often not
  - However, they are easy to use
  - And produce reasonable results

Example MLE

- Ex: observed $\{18, 22, 45, 49, 86\}$
- Assume iid, so take the product of the likihoods
  - $L(\lambda | x_1, x_2, \ldots, x_n) = \prod_i \lambda e^{-\lambda x_i}$
  - Typically, we maximize the log likelihood (LLH) instead:

$$\log\left(\prod_i \lambda e^{-\lambda x_i}\right) = \sum_i \log(\lambda e^{-\lambda x_i}) = \sum_i \log(\lambda) + \log(e^{-\lambda x_i}) = \sum_i \left(\log(\lambda) - \lambda x_i\right)$$

  - Again, this is convex:

$$\begin{aligned} L'(\lambda) &= \sum_i \left(\lambda^{-1} - x_i\right) \\ &= 5\lambda^{-1} - 220 \end{aligned}$$

Why $\log$?

- The math is easier
- It turns products into sums
  - Handy for functions of the form $e^{something}$
- It handles really small numbers well
- The ordering of doesn't change
  - If we have $f(x)$ and we choose an x that maximizes f
  - The same x will maximize $\log(f(x))$

Example MLE

- Setting the derivative equal to zero,

$$\begin{aligned} 0 &= 5\lambda^{-1} - 220 \\ \frac{220}{5} &= \lambda^{-1} \\ \lambda &= \frac{5}{220} = 0.0227 \end{aligned}$$

More Complicated MLE

- Recall from last lecture, we want to predict how many students will turn in their assignments at least 1 hour before the deadline
- $\{18, 22, 45, 49, 86\}$ are the known assignment completion times
- Only 5/10 finished at time 100
- We did this before, right?
- What's a problem we ignored last time?
  - 5 people not done, but they contribute information
  - ? How to model?

More Complicated MLE

- $\{18, 22, 45, 49, 86\}$ are the known assignment completion times
- Only 5/10 finished at time 100
- What's a problem with the last model?
  - 5 people not done, but they contribute information
  - How to model?
- Recall that the exponential distribution is good for modeling arrival times
- It's also has a really nice CDF: $1 - e^{-\lambda x}$
- Each of 5 who have not yet submitted have $x_i \geq 100$
  - So for $i \geq 6$, Pr$[\text{no submission}] = 1 - (1 - e^{-\lambda 100})$
  - Now, $L(\lambda | x_1, x_2, \ldots, x_n) = \prod_{i=1}^{5} (\lambda e^{-\lambda x_i}) \times \prod_{i=6}^{10} e^{-\lambda 100}$
  - ? Where does the first term come from?

More Complicated MLE

- Ex: observed $\{18, 22, 45, 49, 86\}$
- $L(\lambda|.) = \prod_{i=1}^{5} \lambda e^{-\lambda x_i} \times \prod_{i=6}^{10} e^{-\lambda 100}$
- LLH instead: $L(\lambda|.) = \sum_{i=1}^{5} \Big(-\lambda x_i + \log(\lambda)\Big) + \sum_{i=6}^{10} -\lambda 100$
  - Now, minimizing:

$$
\begin{aligned}
L'(\lambda) &= \sum_{i=1}^{5} \left(-x_i + \frac{1}{\lambda}\right) - \sum_{i=6}^{10} 100 \\
&= \sum_{i=1}^{5} \left(-x_i + \frac{1}{\lambda}\right) - 500 \\
&= -220 + \frac{5}{\lambda} - 500 \\
&= \frac{5}{\lambda} - 720
\end{aligned}
$$

More Complicated MLE

- Setting to zero, we have

$$0 = \frac{5}{\lambda} - 720$$
$$720 = \frac{5}{\lambda}$$
$$\lambda = \frac{5}{720}$$

- This is 0.00694
  - So now, probability that a given person (who hasn't turned in) submits within the next 67 hours is 0.372
  - $Pr[\text{submission}] = 1 - e^{-\lambda x} = 1 - * e^{-0.0064 * 167} = 0.372$
  - Before, it was 0.781

Interpreting the Results

- Before, we were more certain that someone who hadn’t submitted the assignment would
- Now we are less certain
- The probability has dropped about in half

Goin' Bayesian

- Complaint regarding MLE approach:
- "It assumes zero knowledge about the parameter(s) you are trying to estimate"
- Do we ever have zero knowledge?
- Bayesians say we always know something
  - Scores so far: $\{99, 92, 94, 94, 88\}$
  - Is the mean best estimated as $(99+92+94+94+88)/5$?
  - Not according to a Bayesian
  - What if I'd never given an assignment with an average > 90 in my career?

Goin' Bayesian

- To a Bayesian:
  - "Learning" is all about updating one's prior opinions in response to evidence
  - It's not about guessing parameters
- "Prior opinions" formally given in the form of a "prior distribution"
  - Pretend I'm a really tough professor
  - The average score on assignments is around 50
  - Highest ever was 65
  - Lowest ever was 35
  - So I choose Normal$(50, 25)$ as the "prior" on the mean assignment score, μ
  - Why variance of 25?
    - Here, 50 is mean, 5 is standard deviation
    - Standard deviation of 5 chosen as 99.7% of mass of Normal is ± 3 std. devs. from mean
    - 50 ± 5 just covers lowest ever, highest ever

Bayes' Rule

- A Bayesian uses data X to update the prior on the parameter set Θ
  - Resulting distribution—$P(\Theta|X)$ is called the "posterior"
- Update is accomplished via "Bayes' Rule"

$$P(\Theta|X) = \frac{P(\Theta)P(X|\Theta)}{P(X)}$$

$$P(\Theta|X) = \frac{\text{Prior on } \Theta \times \text{Likelihood}}{\text{Normalizing constant}}$$

- Can usually drop $P(X)$ as a constant, so we have

$$P(\Theta|X) \propto P(\Theta)P(X|\Theta)$$

Notes on Bayes' Rule

? Why can we drop P(X)?
  - We are asking: What is the posterior distribution, given **fixed** data?
  - Since the data are fixed (observed), P(X) doesn't change the relative ordering
- Change the notation to $\propto$ since there is some constant needed for equality
- P(X) is usually very difficult to compute
- Must integrate over all possible values of the parameters

$$P(X) = \int P(X|\Theta)$$

Bayes' Rule Example

- Scores so far: $\{99, 92, 94, 94, 88\}$
  - Mean score $\mu \sim \text{Normal}(50, 25)$
  - Each score $x_i \sim \text{Normal}(\mu, 16)$
  - Note: 16 chosen arbitrarily (in practice, maybe this is the historical per-score standard deviation)
  - Could also replace 16 with a prior (but don't here, for simplicity)
  - Applying Bayes' rule:

$$P(\mu|\text{data}) \propto \text{Normal}(\mu|50, 25) \prod_i \text{Normal}(x_i|\mu, 16)$$

  - Note: this is a function of 1 variable!

Bayes' Rule Example

$$P(\mu|\text{data})$$

$$\propto \text{Normal}(\mu|50,25)\prod_i \text{Normal}(x_i|\mu,4)$$

$$= 5^{-1}(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}(\mu-50)^2 5^{-2}}\prod_i 4^{-1}(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}(\mu-x_i)^2 4^{-2}} \quad (1)$$

$$\propto e^{-\frac{1}{2}(\mu-50)^2 5^{-2}}\prod_i e^{-\frac{1}{2}(\mu-x_i)^2 4^{-2}} \quad (2)$$

$$= e^{-\frac{1}{2}\left((\mu-50)^2 5^{-2}+\sum_i(\mu-x_i)^2 4^{-2}\right)} \quad (3)$$

$$= e^{-\frac{1}{2}\left(5^{-2}\mu^2-100\times 5^{-2}\mu+2500\times 5^{-2}+\sum_i 4^{-2}\mu^2-2\times 4^{-2}\mu x_i+4^{-2}x_i^2\right)} \quad (4)$$

1. Plug in distributions
2. Drop constants
3. Apply exponent product rule
4. Expand

Bayes' Rule Example

More math...

$$= e^{-\frac{1}{2}\left(5^{-2}\mu^2 - 100\times 5^{-2}\mu + 2500\times 5^{-2} + \sum_i 4^{-2}\mu^2 - 2\times 4^{-2}\mu x_i + 4^{-2}x_i^2\right)} \quad (5)$$

$$\propto e^{-\frac{1}{2}\left(5^{-2}\mu^2 - 4\mu + \sum_i 4^{-2}\mu^2 - 2\times 4^{-2}\mu x_i\right)} \quad (6)$$

$$= e^{\left(2+\frac{1}{16}\sum_i x_i\right)\mu - \left(\frac{1}{50}+\frac{5}{32}\right)\mu^2} \quad (7)$$

$$= e^{a\mu^2 + b\mu} \text{ where } a = -\frac{1}{50} - \frac{5}{32}, b = 2 + \frac{1}{16}\sum_i x_i \quad (8)$$

Now things look relatively simple...

5 From the last slide

6 Simplify & drop constant terms

7 Do some more math

8 Apply exponential quadratic function

Bayes' Rule Example

- We have $P(\mu|\text{data}) \propto e^{a\mu^2+b\mu}$, where:
- $a = -\frac{1}{50} - \frac{5}{32} = -0.17625$
- $b = 2 + \frac{1}{16}\sum_i x_i = 31.1875$
  - By definition, this is $\propto \text{Normal}(-b/(2a), -1/(2a))$
  - Plug in our data to get values for a and b
  - Or, $\text{Normal}(88.475, 2.89)$
- Recall: Questioning if the mean was best estimated as $(99+92+94+94+88)/5 = 93.4$
- Took into account historical data about scores on assignments

Conjugate Priors

- That was a LOT of work!!
- Easier to use a table of conjugate priors
- What is THAT?
  - When you have $\Theta \sim f(\theta_{\text{prior}})$
  - And you have $X \sim g(.)$
  - And you can prove $P(\Theta|X) = f(\Theta|\theta_{\text{post}})$
  - That is, the posterior for Θ is the same family as the prior
  - Then we say f is a "conjugate prior" for g
- There are lots of conjugate priors
- Key tool in Bayesian's toolbox

Conjugate Priors

- Why useful?
- Usually simple rules for computing θ_{post} from X, θ_{prior}
- Ex: Google search “Wikipedia conjugate prior”... first result
- Find row under “continuous distributions”
  - When $g(.)$ (likelihood) is Normal with known variance σ_l^2
  - And $f(\theta_{\text{prior}})$ is Normal(μ_p, σ_p^2)
  - Observed n data points
  - Then posterior is easy!
  - In θ_{post}, we have:

$$\mu = \left(\frac{\mu_p}{\sigma_p^2} + \frac{\sum x_i}{\sigma_l^2}\right) / \left(\frac{1}{\sigma_p^2} + \frac{n}{\sigma_l^2}\right) = \left(\frac{50}{25} + \frac{467}{16}\right) / \left(\frac{1}{25} + \frac{5}{16}\right)$$

$$\sigma^2 = \left(\frac{1}{\sigma_p^2} + \frac{n}{\sigma_l^2}\right)^{-1} = \left(\frac{1}{25} + \frac{5}{16}\right)^{-1}$$

  - Gives the same result, much less fuss!!
  - Often drives choice of distribution

Questions?