COMP 330: Intro to Modeling 2

Chris Jermaine and Kia Teymourian

Rice University

Models Are Parameterized

Normal: μ, σ

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

Exponential: λ

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

Key question: how to choose parameters?

Models Are Parameterized

Normal: μ, σ

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

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
- Item that is, 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
- Probabilistic: Bayesian

Optimization-Based

Goal is to reduce some error metric on example/training data

No direct probabilistic motivation

Example: Least Squares Regression

Ex: I observe $\{18, 22, 45, 49, 86\}$... predict next item?

- Might fit a line to the data
- So $f(t) = m \times t$

Need to choose m

- Might choose least-squares fit
- Eg, choose m to min $l(m) = \sum_i (f(t_i) - x_i)^2$
- $l(m)$ often referred to as a "loss function"

Computing Least-Squares Fit

- This loss function is "convex"
- So just choose unique m where $l'(m) = 0$

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

So loss minimized at $m = 14.96$; next value should be 89.8

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 $\langle \epsilon_1, \epsilon_2, ..., \epsilon_n \rangle$, l_p norm defined as:

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

Common loss functions correspond to various norms:

- l_1 corresponds to mean absolute error
- l_2 to mean squared error/least squares
- l_∞ corresponds to minimax

Maximum Likelihood

Often we have a stochastic model

Ex: observed $\{18, 22, 45, 49, 86\}$

Model is Exponential, unknown λ

How to estimate? Most common: perform MLE

Likelihood

First, need the notion of a "likelihood function"

Best illustrated with an example

- In our case, $f(x_i|\lambda) = \lambda e^{-\lambda x}$
- So $f(x_1, x_2, ..., x_n|\lambda) = \prod_i \lambda e^{-\lambda x_i}$ (iid!)

A "likelihood function" simply turns the parametrization around

- 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

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?

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 "MVUE"
- Under many conditions, error is asymptotically normal

Example MLE

Ex: observed $\{18, 22, 45, 49, 86\}$

- $L(\lambda|x_1, x_2, ..., x_n) = \prod_i \lambda e^{-\lambda x_i}$
- Typically, we maximize the LLH instead:

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

- Again, this is convex:

$$L'(\lambda) = \sum_i x_i + \lambda^{-1}$$
$$= 220 + 5\lambda^{-1}$$

Example MLE

Setting to zero,

$$0 = 220 + 5\lambda^{-1}$$

$$\frac{220}{5} = \lambda^{-1}$$

$$\lambda = \frac{5}{220}$$

More Complicated MLE

Now, imagine $\{18, 22, 45, 49, 86\}$ are assignment completion times

Only 5/10 finished at time 100

What's a problem with the last model?

- 5 people not done contribute info!!
- How to model?

More Complicated MLE

Now, imagine $\{18, 22, 45, 49, 86\}$ are assignment completion times

Only 5/10 finished at time 100

What's a problem with the last model?

- 5 people not done contribute info!!
- How to model?

Each of 5 who have not yet arrived have $x_i \geq 100$

- CDF of exponential is $1 - e^{\lambda x}$
- So for $i \geq 5$, Pr[**no submission**] $= 1 - (1 - e^{-\lambda 100})$
- Now, $L(\lambda | x_1, x_2, ..., x_n) = \prod_{i=1}^{5} \lambda e^{-\lambda x_i} \times \prod_{i=6}^{10} e^{-\lambda 100}$

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: $L(\lambda|.) = \sum_{i=1}^{5} -\lambda x_i + log(\lambda) + \sum_{i=6}^{10} -\lambda 100$

- Now, minimizing:

$$L'(\lambda) = -\sum_{i=1}^{5} x_i + \frac{1}{\lambda} - \sum_{i=6}^{10} 100$$

$$= \frac{5}{\lambda} - 500 - \sum_{i=1}^{5} x_i$$

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

More Complicated MLE

Setting to zero, we have

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

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

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

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?

- Scores so far: $\{99, 92, 94, 94, 88\}$
- Is mean best estimated as $(99 + 92 + 94 + 94 + 88)/5$?
- What if I'd never had an assignment with avg > 90 in my life?

Goin' Bayesian

To a Bayesian:

- "Learning" is all about updating one's prior opinions in response to evidence

"Prior opinions" formally given in the form of a "prior distribution"

- Pretend I'm really nasty :-)
- My average assignment score is around 50
- Highest ever was 70
- Lowest ever was 30
- So I choose Normal$(\mathbf{50}, \mathbf{10})$ as the "prior" on the mean assignment score μ

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"

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

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

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

Bayes' Rule Example

Scores so far: $\{99, 92, 94, 94, 88\}$

- Mean score $\mu \sim \textbf{Normal}(50, 5)$
- Each score $x_i \sim \textbf{Normal}(\mu, 4)$
- Applying Bayes' rule:

$$P(\mu|\text{data}) \propto \text{Normal}(\mu|50, 10) \prod_i \text{Normal}(x_i|\mu, 4)$$

Bayes' Rule Example

Lots o' math!! No!! $P(\mu|\textbf{data})$

$$\propto \textbf{Normal}(\mu|50,5)\prod_i \textbf{Normal}(x_i|\mu,4)$$

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

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

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

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

Come up for air...

Bayes' Rule Example

More math...

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

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

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

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

Now things look quite simple...

Bayes' Rule Example

We have $P(\mu|\textbf{data}) \propto e^{a\mu^2+b\mu}$, where:

$$a = -\tfrac{1}{50} - \tfrac{5}{32} = -0.17625, b = 2 + \tfrac{1}{16}\textstyle\sum_i x_i = 31.1875$$

- By definition, this is $\propto$ **Normal**$(-b/(2a), \sqrt{-1/(2a)})$
- Or, **Normal**$(88.475, 1.7)$

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

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 σ
- And $f(\theta_{\text{prior}})$ is Normal(μ_0, σ_0)
- Then posterior is easy!
- In θ_{post}, we have:

$$\mu = \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum x_i}{\sigma^2}\right) / \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) = \left(\frac{50}{25} + \frac{467}{16}\right) / \left(\frac{1}{25} + \frac{5}{16}\right)$$

$$\sigma^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} = \left(\frac{1}{25} + \frac{5}{16}\right)^{-1}$$

- Gives the same result, much less fuss!!

What Do Bayesians Do When Things Are Nasty?

That is, no conjugate prior in table...

Can't easily do the math...

Do they just give up??

- Not so easily!
- More on this later...

Questions?