MET CS 777 - Big Data Analysis

Module 3:
Probabilistic Methods

Dimitar Trajanov

Table of contents

What is a Model?

What is a Model?

We can find many different definitions.

- Traditional statistical definition:
  "A set of assumptions regarding the stochastic process that generated the data"
  Here we assume that a stochastic process generated the data.
  We would like to learn from data and find out what the stochastic process was which generated the data.
  A stochastic or random process is a mathematical object usually defined as a family of random variables and it changes randomly over time.
- More modern definition:
  "An algorithm that can be used to generate an artifact explaining the data" "A mathematical object that helps us to understand better past and present, and be able to use it for predicting the future."

What's the difference?

Why Do We need Models?

- Real data are big, complex, difficult to understand
- A model is (hopefully!) compact, simple and comprehensible so that it can help us to understand the data.
- We want to make some assumptions and do simplification.

Just as important:

- Models can often be used to make predictions of future events.
- Models can be use to understand the data patterns and relations.

Many (not all!) models rely on the idea of probability

- ”The extent to which an event is likely to occur, measured by the ratio of the favorable cases to the whole number of cases possible”

What about infinitely many possible events?

Then probability tends to zero

- Example: The chance I jump exactly 3 feet
- Example: Chance class ends at exactly 8:45 pm.
- Example: The chance it takes 5 hours to complete Assignment-2

Probability Density

Probability density is around this problem

- Measures the relative likelihood of an event - not absolute values

Probability of Assignment 2 to take 5 hours - nonsensical

But ...

- Likelihood that the Assignment-2 takes 5 hours is 5 times than it takes 1 hour
- So better, Sensical!

Probability Density Function

A PDF is a function that computes the relative probability of an event.

For example:

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

Most famous: normal PDF

- A PDF can be used to calculate the probability of a range of events
- $\int_a^b f(x)dx$ is the probability we see a value in range a to b

Choosing a Model

We can see that all models are wrong but they are very useful to describe things.

Remember:

- A model is (hopefully!) compact, simple and comprehensible
- We choose models to reduce, simplify and comprehend data
- Hopefully, without incurring inaccuracy!!

Data Modeling

We have 4 main tasks

- **Choosing the model** - choose main concepts, family, complexity, hyperparameters
- **Learning the model** - Fit the model to data by learning the hyperparameters from data
- **Validate the model** - Check if the model match the requirements
- **Applying the model** - Actual usage of the data model to understand and explain the data, or predict data (Future data or past data)

Example Model

Example: Predicting Grade In Class

student has completed 5 out of 10 assignments.

- **Want to predict grade in class**

First Step, **choose a model**:

- For example assume $X_i \sim Normal(\mu, \sigma)$ (Why normal?)
- i is the identity of the assignment (i_{th} assignment)
- Note: X_i is a random variable controlling grade
- $f_{X_i}(x)$ gives relative likelihood X_i takes value x
- (or probability if X_i is discrete!)
- So that $f_{X_i}(x) = f_{Normal}(x|\mu, \sigma)$

Example: Predicting Grade In Class

Let assume that a student got scores of {89, 92, 78, 94, 88}

- Estimate mean $\mu = 88.2, \sigma^2 = 30.56$
- Thus $\sum_{i=6\ldots10} X_i \sim Normal(88.2 \times 5, (30.56 \times 5)^{0.5})$
- 95% confidence on sum: $882 \pm 2 \times 12.36$
- Or, 95% confidence on average: 88.2 ± 2.47

Another Example: Assignment Turn In

5 out of 10 students have completed the assignment

Turn In Deadline is 168 hours (one week) to complete the assignment

We want to predict how many have completed by 1 hour before due date

How should we model this?

Another Example: Assignment Turn In

5 out of 10 students have completed the assignment

Turn In Deadline is 168 hours (one week) to complete the assignment

We want to predict how many have completed by 1 hour before due date

How should we model this?

We have a probability and two possible outcomes (Turned in by deadline or Not turned in by deadline)

Looks like Binomial distribution

Binomial distribution

It is a discrete distribution

- Has 2 parameters
  n = number of independent experiments
  p = probability of success
- The probability of getting exactly k successes in n trials is given by the probability mass function:

  Probability Mass Function (PMF) $= \binom{n}{k} p^k (1-p)^{n-k}$
- Mean: np
- Variance: $np(1-p)$
- Good for modeling Yes/No choices, n times
- All of the trials must be independent
- Degenerative form is the Bernoulli distribution, when n = 1

Another Example: Assignment Turn In

5 out of 10 students have completed the assignment

Turn In Deadline is 168 hours (one week) to complete the assignment

We want to predict how many have completed by 1 hour before due date

How should we model this?

Another Example: Assignment Turn In

5 out of 10 students have completed the assignment

Turn In Deadline is 168 hours (one week) to complete the assignment

We want to predict how many have completed by 1 hour before due date

- X_i : number of hours after assignment student i turns in
- Assume $X_i \sim Exponential(\lambda)$

Another Example: Assignment Turn In

5 out of 10 students have completed the assignment

Turn In Deadline is 168 hours (one week) to complete the assignment

We want to predict how many have completed by 1 hour before due date

- X_i : number of hours after assignment student i turns in
- Assume $X_i \sim Exponential(\lambda)$
- Exponential: $f_{Exp}(x|\lambda) = \lambda e^{-\lambda x}$
- Memoryless property!
- It means if we waited unites so far ...
- $f_{Exp}(x|\lambda, x \geq t) = f_{Exp}(x - t|\lambda)$

Another Example: Assignment Turn In

Times so far at time tick 100 are : $\{18, 22, 45, 49, 86\}$

- We know mean of exponential is λ^{-1}
- In our data case mean is, $41 = \lambda^{-1}$ so $\lambda \approx 0.0227$
- Look up the Cumulative Distribution Function (CDF): $1 - e^{-\lambda x}$
- Is 0.878 at $167 - 100$
- So prob of each remaining person turning in by deadline is 0.781

Another Example: Assignment Turn In

Times so far at time tick 100 are : $\{18, 22, 45, 49, 86\}$

- We know mean of exponential is λ^{-1}
- In our data case mean is, $41 = \lambda^{-1}$ so $\lambda \approx 0.0227$
- Look up the Cumulative Distribution Function (CDF): $1 - e^{-\lambda x}$
- Is 0.878 at $167 - 100$
- So prob of each remaining person turning in by deadline is 0.781
- What about number of people?

Another Example: Assignment Turn In

5 people, each with 0.781 chance of turning in at deadline -1

How to model this?

Another Example: Assignment Turn In

5 people, each with 0.781 chance of turning in at deadline -1

How to model this?

- $N \sim Binomial(0.878, 5)$
- N is the number turning in
- $Pr(N = 5) = 0.291$ = prob all 10 turn in
- $Pr(N = 4) = 0.698$ = prob 9+ turn in
- $Pr(N = 3) = 0.926$ = prob 8+ turn in
- $Pr(N < 3) = 0.074$ = prob < 8 turn in

Probabilistic: Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE)

Often we have a stochastic model

Example: 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 that, $f(x_1, x_2, ...x_n|\lambda) = \prod_{n=1} e^{\lambda x_i}$ (iid!)

A "likelihood function" simply turns the parameterization around.

- So $L(\lambda|x_1, x_2, ...x_n) = \prod_{n=1} e^{\lambda x_i}$
- 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 parameters, 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: This is closely related to least squares!!

Why do we like this?

- Under many conditions, it is the minimum-variance unbiased estimator (MVUE)
- Under many conditions, error is asymptotically normal

Example MLE

Example, we observed {18, 22, 45, 49, 86}

- So $L(\lambda|x_1, x_2, ...x_n) = \prod_{n=1}^{} \lambda e^{-\lambda x_i}$
- Typically, we maximize the log-likelihood (LLH) instead:

$$\sum_{i=1} log(\lambda e^{-\lambda x_i}) = \sum_{i=1} -\lambda x_i + log(\lambda)$$

- Again, this is convex:

derivative = zero, to find the minimum

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

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

More Complicated MLE

Example $\{18, 22, 45, 49, 86\}$

$L(\lambda|\cdot) = \prod_{i=1}^{5} \lambda e^{-\lambda x_i} \times \prod_{i=6}^{10} e^{-\lambda 100}$

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

- ▷ Now, minimizing:

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

More Complicated MLE

Setting to zero, we have

$$
\begin{aligned}
0 &= \frac{5}{\lambda} - 720 \\
720 &= \frac{5}{\lambda} \\
\lambda &= \frac{5}{720}
\end{aligned}
$$

Probabilistic: Bayesian

Complaint regarding MLE approach:

"It assumes zero knowledge about the parameter(s) you are trying to estimate, right? :-("

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?

We have prior knowledge!

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 5
- ▷ Highest ever was 70
- ▷ Lowest ever was 3
- ▷ So I choose Normal(50, 5) 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 $\mu \sim Normal(50, 5)$
- Each score $x_i \sim Normal(\mu, 4)$
- Applying Bayes' rule:

$$P(\mu|data) \propto Normal(\mu|50, 5) \prod_{i=1} Normal(x_i|\mu, 4)$$

Bayes' Rule Example

Lots do some math here!! No!!... $P(\mu|data)$

$$P(\mu|data) \propto Normal(\mu|50,5) \prod_i Normal(x_i|\mu,4)$$

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

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

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

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

Bayes' Rule Example

A bit more Math ...

$$
\begin{aligned}
&= e^{-\frac{1}{2}\left(5^{-2}\mu^2 - 100\times 5^{-2}\mu + 2500\times 5^{-2} + \sum_i 4^{-2}\mu^2 - 2\times 4^{-2}\mu\times x_i + 4^{-2}{x_i}^2\right)} \\
&\propto e^{-\frac{1}{2}\left(5^{-2}\mu^2 - 4\mu + \sum_i 4^2\mu^2 - 2\times 4^{-2}\mu\times x_i\right)} \\
&= e^{\left(2+\frac{1}{16}\sum_i x_i\right)\mu - \left(\frac{1}{50}+\frac{5}{32}\right)\mu^2} \\
&= e^{a\mu^2 + b\mu}
\end{aligned}
$$

Where

$$
\begin{aligned}
a &= -\frac{1}{50} - \frac{5}{32} \\
b &= 2 + \frac{1}{16}\sum_i x_i
\end{aligned}
$$

Now this looks better and quite simple ...

Bayes' Rule Example

We have

$$P(\mu|data) \propto e^{a\mu^2+b\mu}$$

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 $\propto Normal(-b/(2a), \sqrt{-1/(2a)})$**
- **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 $\Theta \sim f(\theta_{prior})$
- And you have $X \sim g(.)$
- And you can prove $P(\Theta|X) = f(\Theta|\theta_{post})$
- 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}

Google search "Wikipedia conjugate prior"... first result

Find row under "Continuous Distributions"

- **When $g(.)$ (likelihood) is Normal with known σ**
- **And $f(\theta_{prior})$ is $Normal(\mu_0, \sigma_0)$**
- **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 work!!