Parametric Models: from data to models

Pradeep Ravikumar
Co-instructor: Manuela Veloso

Machine Learning 10-701
Jan 22, 2018

Recall: Model-based ML

- Learning: From data to model
  - A model explains how the data was "generated"
  - **E.g. given (symptoms, diseases) data, a model explains which symptoms arise from which diseases**

- Inference: From model to knowledge
  - Given the model, we can then answer questions relevant to us
  - **E.g. given (symptom, disease) model, given some symptoms, what is the disease?**

Model Learning: Data to Model

- What are some general principles in going from data to model?
- What are the guarantees of these methods?

LET US CONSIDER THE EXAMPLE OF A SIMPLE MODEL

Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have a coin, if I flip it, what's the probability it will fall with the head up?
  - You say: Please flip it a few times:

Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have a coin, if I flip it, what's the probability it will fall with the head up?
  - You say: Please flip it a few times:

  - You say: The probability is **3/5**
  - **He says: Why???**
  - You say: Because… frequency of heads in all flips

Questions

- Why frequency of heads?

- How good is this estimation?
  - Would you be willing to bet money on your guess of the probability?
  - Why not?

Model-based Approach

- First we need a model that would explain the experimental data
- What is the experimental data?
- Coin Flips

Model

- First we need a model that would capture the experimental data
- What is the experimental data?
- Coin Flips

Model

- A model for coin flips
  - Bernoulli Distribution
- X is a random variable with Bernoulli distribution when:
  - X takes values in {0,1}
  - $P(X = 1) = p$
  - $P(X = 0) = 1 - p$
  - Where p in [0,1]

Model

- X is a random variable with Bernoulli distribution when:
  - X takes values in {0,1}
  - $P(X = 1) = p$
  - $P(X = 0) = 1 - p$
  - Where p in [0,1]
- X = 1 i.e. heads with probability p, and X = 0 i.e. tails with probability 1 – p
  - Coin with probability of flipping heads = p
- And we draw **independent** samples that are **identically distributed** from same distribution
  - flip the same coin multiple times

Bernoulli distribution

Data, D = 

- P(Heads) = θ, P(Tails) = 1-θ

- Flips are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Bernoulli distribution

<u>Choose θ that maximizes the probability of observed data</u>

Probability of one coin flip

Let's say we observe a coin flip $X \in \{0, 1\}$.

The probability of this coin flip,
given a Bernoulli distribution with parameter p:

$$p^X (1 - p)^{1-X}.$$

Equal to p when $X = 1$, and equal to $(1 - p)$ when $X = 0$.

Probability of Multiple Coin Flips

$$\text{Probability of Data} = \mathbb{P}(X_1, X_2, \ldots, X_n; \theta)$$

Probability of Multiple Coin Flips

$$\begin{aligned}\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\ &= P(X_1)\, P(X_2) \ldots P(X_n)\end{aligned}$$

…Independence of samples

Probability of Multiple Coin Flips

$$\begin{aligned}\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\ &= P(X_1)\, P(X_2) \ldots P(X_n) \\ &= \prod_{i=1}^{n} P(X_i)\end{aligned}$$

Probability of Multiple Coin Flips

$$\begin{aligned}\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\ &= P(X_1)\, P(X_2) \ldots P(X_n) \\ &= \prod_{i=1}^{n} P(X_i) \\ &= \prod_{i=1}^{n} p^{X_i}\, (1-p)^{1-X_i}\end{aligned}$$

...probability of a Bernoulli sample

Probability of Multiple Coin Flips

$$\begin{aligned}\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\ &= P(X_1)\, P(X_2) \ldots P(X_n) \\ &= \prod_{i=1}^{n} P(X_i) \\ &= \prod_{i=1}^{n} p^{X_i}\, (1-p)^{1-X_i} \\ &= p^{\sum_{i=1}^{n} X_i}\, (1-p)^{n-\sum_{i=1}^{n} X_i}\end{aligned}$$

$$\ldots\ p^a\, p^b = p^{a+b}$$

Probability of Multiple Coin Flips

$$
\begin{aligned}
\text{Probability of Data} &= \mathbb{P}(X_1, X_2, \ldots, X_n; \theta) \\
&= P(X_1)\, P(X_2) \ldots P(X_n) \\
&= \prod_{i=1}^{n} P(X_i) \\
&= \prod_{i=1}^{n} p^{X_i}\, (1-p)^{1-X_i} \\
&= p^{\sum_{i=1}^{n} X_i}\, (1-p)^{n-\sum_{i=1}^{n} X_i} \\
&= p^{n_h}\, (1-p)^{n-n_h}.
\end{aligned}
$$

where n_h is the number of heads,
n is the total number of coin flips

Maximum Likelihood Estimator (MLE)

The MLE solution is then given by solving the following problem:

$$\begin{aligned}\widehat{p} &= \arg\max_{p} \mathbb{P}(X_1, \ldots, X_n; p) \\ &= \arg\max_{p} \left\{p^{n_h} (1-p)^{n-n_h}\right\}\end{aligned}$$

Maximum Likelihood Estimator (MLE)

The MLE solution is then given by solving the following problem:

$$
\begin{aligned}
\widehat{p} &= \arg\max_{p} \mathbb{P}(X_1, \ldots, X_n; p) \\
&= \arg\max_{p} \left\{ p^{n_h} (1-p)^{n-n_h} \right\} \\
&= \arg\max_{p} \left\{ n_h \log p + (n - n_h) \log(1-p) \right\}
\end{aligned}
$$

…argmax_x f(x) = argmax_x log f(x)

MLE for coin flips

The MLE solution is then given by solving the following problem:

$$
\begin{aligned}
\widehat{p} &= \arg\max_{p} \left\{ n_h \log p + (n - n_h)\, \log(1-p) \right\} \\
\Longrightarrow \frac{n_h}{\widehat{p}} &- \frac{n-n_h}{1-\widehat{p}} = 0 \\
\Longrightarrow \widehat{p} &= \frac{n_h}{n}.
\end{aligned}
$$

Maximum Likelihood Estimation

Choose θ that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D \mid \theta)$$

MLE of probability of head:

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$ **= 3/5**

“Frequency of heads”

How many flips do I need?

$$\widehat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: θ = 3/5, it is the MLE!
- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, it is the MLE!
- **He says: If you get the same answer, would you prefer to flip 5 times or 50 times?**
- You say: Hmm… The more the merrier???
- He says: Is this why I am paying you the big bucks???

SO FAR:
THE MLE IS A CLASS OF ESTIMATORS THAT ESTIMATE MODEL FROM DATA

KEY QUESTION: HOW GOOD IS THE MLE (OR ANY OTHER ESTIMATOR)?

How good is this MLE: Infinite Sample Limit

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

How good is this MLE: Infinite Sample Limit

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the "true" coin flip probability p.

How good is this MLE: Infinite Sample Limit

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the "true" coin flip probability p.

More formally: as we flip more and more times, we want our estimator to converge (in probability) to the true coin flip probability.

How good is this MLE: Infinite Sample Limit

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the “true” coin flip probability p.

More formally: as we flip more and more times, we want our estimator to converge (in probability) to the true coin flip probability.

This property is known as **consistency**.

How good is this MLE: Infinite Sample Limit

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the "true" coin flip probability p.

More formally: as we flip more and more times, we want our estimator to converge (in probability) to the true coin flip probability.

This property is known as **consistency**.

Do we get that $\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i \to p$ in probability as $n \to \infty$?

How good is this MLE: Infinite Sample Limit

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the "true" coin flip probability p.

More formally: as we flip more and more times, we want our estimator to converge (in probability) to the true coin flip probability.

This property is known as **consistency**.

Do we get that $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \to p$ in probability as $n \to \infty$?

By the Law of Large Numbers!

How good is this MLE: Infinite Sample Limit

If we flipped the coin infinitely many times, and then computed our estimator, what would it look like?

It would be great if it would then be equal to the "true" coin flip probability p.

More formally: as we flip more and more times, we want our estimator to converge (in probability) to the true coin flip probability.

This property is known as **consistency**.

Do we get that $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \to p$ in probability as $n \to \infty$?

By the Law of Large Numbers!

…since the sample mean converges to E(X) = p

How good is this MLE: Infinite Trial Average

If we repeated this experiment infinitely many times, i.e. flip a coin n times and calculate our estimator, and then took an average of our estimator over the infinitely many trials.

What would the average look like?

How good is this MLE: Infinite Trial Average

If we repeated this experiment infinitely many times, i.e. flip a coin n times and calculate our estimator, and then took an average of our estimator over the infinitely many trials.

What would the average look like?

Formally: the estimator $\widehat{p}$ is random: it depends on the samples (i.e. coin flips) drawn from a Bernoulli distribution with parameter p.

What would the expectation of the estimator be?

How good is this MLE: Infinite Trial Average

If we repeated this experiment infinitely many times, i.e. flip a coin n times and calculate our estimator, and then took an average of our estimator over the infinitely many trials.

What would the average look like?

Formally: the estimator $\widehat{p}$ is random: it depends on the samples (i.e. coin flips) drawn from a Bernoulli distribution with parameter p.

What would the expectation of the estimator be?

It would be great if this expectation be equal to the "true" coin flip probability.

How good is this MLE: Infinite Trial Average

If we repeated this experiment infinitely many times, i.e. flip a coin n times and calculate our estimator, and then took an average of our estimator over the infinitely many trials.

What would the average look like?

Formally: the estimator $\widehat{p}$ is random: it depends on the samples (i.e. coin flips) drawn from a Bernoulli distribution with parameter p.

What would the expectation of the estimator be?

It would be great if this expectation be equal to the "true" coin flip probability.

This property is called **unbiasedness**.

How good is this MLE?

It would be great if this expectation be equal to the "true" coin flip probability.

This property is called **unbiasedness**.

$$\begin{aligned} \mathbb{E}(\widehat{p}) &= \mathbb{E}\left(\frac{n_h}{n}\right) \\ &= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \end{aligned}$$

How good is this MLE?

It would be great if this expectation be equal to the "true" coin flip probability.

This property is called **unbiasedness**.

$$\begin{aligned}
\mathbb{E}(\widehat{p}) &= \mathbb{E}\left(\frac{n_h}{n}\right) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i)
\end{aligned}$$

…linearity of expectation:
E(a X + b Y) = a E(X) + b E(Y)

How good is this MLE?

It would be great if this expectation be equal to the “true” coin flip probability.

This property is called **unbiasedness**.

$$
\begin{aligned}
\mathbb{E}(\widehat{p}) &= \mathbb{E}\left(\frac{n_h}{n}\right) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \\
&= \mathbb{E}X_1
\end{aligned}
$$

How good is this MLE?

It would be great if this expectation be equal to the “true” coin flip probability.

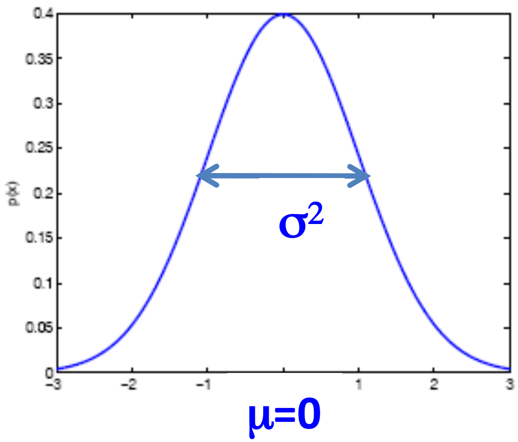
This property is called **unbiasedness**.

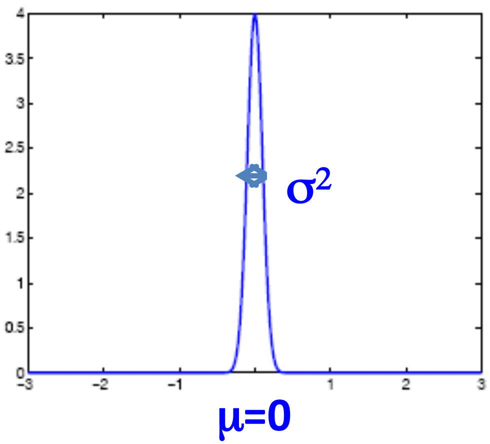
$$
\begin{aligned}
\mathbb{E}(\widehat{p}) &= \mathbb{E}\left(\frac{n_h}{n}\right) \\
&= \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \\
&= \mathbb{E}X_1 \\
&= p.
\end{aligned}
$$

What about continuous variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- **You say: Let me tell you about Gaussians…**

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad = N(\mu,\sigma^2)$$





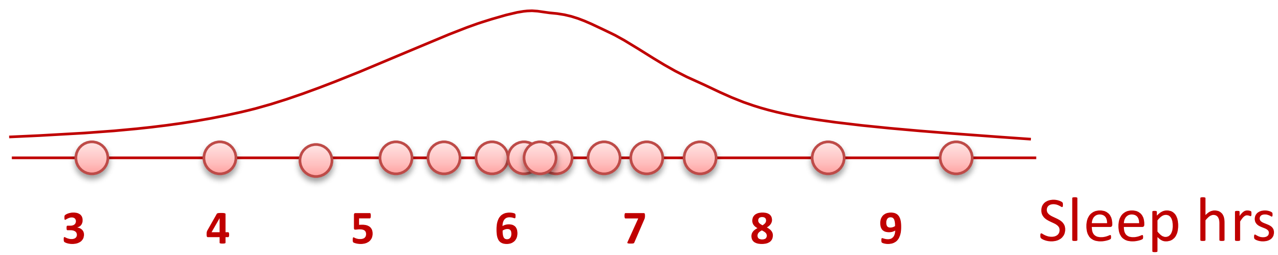
Properties of Gaussians

- affine transformation (multiplying by scalar and adding a constant)
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b$! $Y \sim N(a\mu+b,a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y$! $Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

Gaussian distribution

Data, D =



- Parameters: μ – mean, σ^2 - variance
- Sleep hrs are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Gaussian distribution

MLE for Gaussian mean and variance

$$\widehat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

Note: MLE for the variance of a Gaussian is **biased**

– Expected result of estimation is **not** true parameter!

– Unbiased variance estimator:

$$\widehat{\sigma}^2_{unbiased} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \widehat{\mu})^2$$

MLE for parametric models

Data: $X_1, X_2, \ldots, X_n$.

Model: $P(X;\theta)$ with parameters θ.

MLE for parametric models

Data: $X_1, X_2, \ldots, X_n$.

Model: $P(X; \theta)$ with parameters θ.

Assumption: Data drawn $i.i.d$ from distribution $P(X; \theta^*)$ for some unknown θ^*.

MLE for parametric models

Data: $X_1, X_2, \ldots, X_n$.

Model: $P(X;\theta)$ with parameters θ.

Assumption: Data drawn $i.i.d$ from distribution $P(X;\theta^*)$ for some unknown θ^*.

Mission (should you choose to accept it): recover θ^* from data $X_1, X_2, \ldots, X_n$.

MLE for parametric models

Data: $X_1, X_2, \ldots, X_n$.

Model: $P(X;\theta)$ with parameters θ.

Assumption: Data drawn $i.i.d$ from distribution $P(X;\theta^*)$ for some unknown θ^*.

Mission (should you choose to accept it): recover θ^* from data $X_1, X_2, \ldots, X_n$.

R. A. Fisher

MLE for parametric models

Data: $X_1, X_2, \ldots, X_n$.

Model: $P(X;\theta)$ with parameters θ.

Assumption: Data drawn $i.i.d$ from distribution $P(X;\theta^*)$ for some unknown θ^*.

Mission (should you choose to accept it): recover θ^* from data $X_1, X_2, \ldots, X_n$.

Likelihood Function: $L(\theta) := \prod_{i=1}^n P(X_i;\theta)$
The probability of seeing data $X_1, X_2, \ldots, X_n$ assuming parameters were θ.

MLE for parametric models

Data: $X_1, X_2, \ldots, X_n$.

Model: $P(X;\theta)$ with parameters θ.

Assumption: Data drawn $i.i.d$ from distribution $P(X;\theta^*)$ for some unknown θ^*.

Mission (should you choose to accept it): recover θ^* from data $X_1, X_2, \ldots, X_n$.

Likelihood Function: $L(\theta) := \prod_{i=1}^{n} P(X_i;\theta)$
The probability of seeing data $X_1, X_2, \ldots, X_n$ assuming parameters were θ.

Maximum Likelihood Estimator (MLE): find that parameter θ that would maximize the likelihood of θ

MLE for parametric models

Data: $X_1, X_2, \ldots, X_n$.

Model: $P(X;\theta)$ with parameters θ.

Assumption: Data drawn $i.i.d$ from distribution $P(X;\theta^*)$ for some unknown θ^*.

Mission (should you choose to accept it): recover θ^* from data $X_1, X_2, \ldots, X_n$.

Likelihood Function: $L(\theta) := \prod_{i=1}^n P(X_i;\theta)$
The probability of seeing data $X_1, X_2, \ldots, X_n$ assuming parameters were θ.

Maximum Likelihood Estimator (MLE): find that parameter θ that would maximize the likelihood of θ

i.e. pick the θ that would maximize the probability of having seen the data that we do see

Unbiasedness

An estimator $\widehat{\theta}(X_1, \ldots, X_n)$ where $X_i \sim P(X; \theta^*)$ is unbiased if

$$\mathbb{E}(\widehat{\theta}) = \theta^*.$$

MLE is "asymptotically" unbiased i.e. there are some error terms that go to zero as a function of n, the number of samples

Consistency

An estimator $\widehat{\theta}(X_1, \ldots, X_n)$ where $X_i \sim P(X; \theta^*)$ is consistent if $\widehat{\theta} \to \theta^*$ in probability as $n \to \infty$.

MLE is consistent under some mild regularity conditions on the model, and when the model size is fixed.

How many flips?

- But recall the Billionaire's question:
  - How many flips would you prefer: 5 or 50?
  - How many flips would you need to be willing to bet money on your answer?
- Unbiasedness and Consistency do not answer this question
- We need convergence rates for our estimator

Simple bound (Hoeffding's inequality)

- For $n = \alpha_H + \alpha_T$, and $\widehat{\theta}_{MLE} = \dfrac{\alpha_H}{\alpha_H + \alpha_T}$

- Let θ^* be the true parameter, for any $\varepsilon > 0$:

$$P(|\,\widehat{\theta} - \theta^*\,| \geq \epsilon) \;\leq\; 2e^{-2n\epsilon^2}$$

PAC Learning

- PAC: Probably Approximate Correct
- Billionaire says: I want to know the coin parameter θ, within ε = 0.1, with probability at least 1-δ = 0.95. How many flips?

$$P(|\,\widehat{\theta} - \theta^*\,| \geq \epsilon) \;\; \leq \;\; 2e^{-2n\epsilon^2}$$

PAC Learning

- PAC: Probably Approximate Correct
- Billionaire says: I want to know the coin parameter θ, within ε = 0.1, with probability at least 1-δ = 0.95. How many flips?

$$P(\mid \widehat{\theta} - \theta^* \mid \geq \epsilon) \quad \leq \quad 2e^{-2n\epsilon^2}$$

Suffice to have n large enough for RHS to be less than \delta

PAC Learning

- PAC: Probably Approximate Correct
- Billionaire says: I want to know the coin parameter θ, within ε = 0.1, with probability at least 1-δ = 0.95. How many flips?

$$P(|\,\widehat{\theta} - \theta^*\,| \geq \epsilon) \quad \leq \quad 2e^{-2n\epsilon^2}$$

$$\begin{aligned} 2e^{-2n\epsilon^2} &< \delta \\ -2n\epsilon^2 &< \ln(\delta/2) \\ 2n\epsilon^2 &> \ln(2/\delta) \\ n &> \frac{\ln(2/\delta)}{2\epsilon^2} \end{aligned}$$

PAC Learning

- PAC: Probably Approximate Correct
- Billionaire says: I want to know the coin parameter θ, within $\varepsilon = 0.1$, with probability at least $1-\delta = 0.95$.
  How many flips?

$$P(|\,\hat{\theta} - \theta^*\,| \geq \epsilon) \;\; \leq \;\; 2e^{-2n\epsilon^2}$$

Sample complexity

$$n \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$

From data to model

- Well-studied question in Statistics
  - Estimators e.g. MLE
  - Guarantees (consistency, unbiasedness, convergence rates)
- What has Machine Learning contributed to this statistical question:
  - Specific kinds of guarantees e.g. sample complexity
  - New tools to derive guarantees (VC Dimension, etc.)
  - Computational Issues

Computational Issues

- MLE

$$\max_{\theta} \prod_{i=1}^{n} P(X_i; \theta)$$

$$\max_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log P(X_i; \theta)$$

- Maximizing the log of likelihood easier computationally than maximizing likelihood
- These two optimization problems have the same optima

Computational Issues

- When number of parameters, or number of samples n is large, computing the MLE is a **large-scale optimization problem**
- Well-studied problem in optimization/operations research
- Machine Learning has contributed considerably via:
  – Better understanding of optimization problems that arise from statistical estimators such as MLE (in contrast to general optimization problems)