

10-601 Introduction to Machine Learning

Machine Learning Department School of Computer Science Carnegie Mellon University

MLE/MAP + Naïve Bayes

Matt Gormley Lecture 17 Mar. 20, 2020

Reminders

- Homework 5: Neural Networks
 - Out: Fri, Feb 28
 - Due: Sun, Mar 22 at 11:59pm
- Homework 6: Learning Theory / Generative Models
 - Out: Fri, Mar 20
 - Due: Fri, Mar 27 at 11:59pm

TIP: Do the readings!

- Today's In-Class Poll
 - http://poll.mlcourse.org
- Matt's new after-class office hours (on Zoom)

MLE AND MAP

Likelihood Function

One R.V.

• Suppose we have N samples D = $\{x^{(1)}, x^{(2)}, ..., x^{(N)}\}$ from a random variable X

```
• The likelihood function:

- Case 1: X is discrete with pmf p(x|\theta)

L(\theta) = p(x^{(1)}|\theta) p(x^{(2)}|\theta) \dots p(x^{(N)}|\theta)

- Case 2: X is continuous with pdf f(x|\theta)

L(\theta) = f(x^{(1)}|\theta) f(x^{(2)}|\theta) \dots f(x^{(N)}|\theta)
```

- The log-likelihood function:
 - Case 1: X is **discrete** with pmf $p(x|\theta)$ $\ell(\theta) = \log p(x^{(1)}|\theta) + ... + \log p(x^{(N)}|\theta)$
 - Case 2: X is **continuous** with pdf $f(x|\theta)$ $\ell(\theta) = \log f(x^{(1)}|\theta) + ... + \log f(x^{(N)}|\theta)$

Likelihood Function

Two R.V.s

Suppose we have N **samples** D = $\{(x^{(1)}, y^{(1)}), ..., (x^{(N)}, y^{(N)})\}$ from a pair of random variables X, Y

The **conditional likelihood** function:

- Case 1: Y is **discrete** with pmf $p(y \mid x, \theta)$ + Log Peg. $L(\theta) = p(y^{(1)} \mid x^{(1)}, \theta) \dots p(y^{(N)} \mid x^{(N)}, \theta)$ Case 2: Y is **continuous** with pdf $f(y \mid x, \theta)$ + Line Peg. $L(\theta) = f(y^{(1)} \mid x^{(1)}, \theta) \dots f(y^{(N)} \mid x^{(N)}, \theta)$

The **joint likelihood** function:

- Case 1: X and Y are **discrete** with pmf $p(x,y|\theta)$ $L(\theta) = p(x^{(1)}, y^{(1)}|\theta) \dots p(x^{(N)}, y^{(N)}|\theta)$
- Case 2: X and Y are **continuous** with pdf $f(x,y|\theta)$ $L(\theta) = f(x^{(1)}, y^{(1)}|\theta) \dots f(x^{(N)}, y^{(N)}|\theta)$

Likelihood Function

Two R.V.s

- Suppose we have N samples D = $\{(x^{(1)}, y^{(1)}), ..., (x^{(N)}, y^{(N)})\}$ from a pair of random variables X, Y
- The joint likelihood function:
 - Case 1: X and Y are **discrete** with pmf $p(x,y|\theta)$ $L(\theta) = p(x^{(1)}, y^{(1)}|\theta) \dots p(x^{(N)}, y^{(N)}|\theta)$
 - Case 2: X and Y are **continuous** with pdf $f(x,y|\theta)$ $L(\theta) = f(x^{(1)}, y^{(1)}|\theta) \dots f(x^{(N)}, y^{(N)}|\theta)$

Mixed discrete/continuous!

- Case 3: Y is **discrete** with pmf $p(y|\beta)$ and X is **continuous** with pdf $f(x|y,\alpha)$ $L(\alpha,\beta) = f(x^{(1)}|y^{(1)},\alpha) p(y^{(1)}|\beta) \dots f(x^{(N)}|y^{(N)},\alpha) p(y^{(N)}|\beta)$
- Case 4: Y is **continuous** with pdf $f(y|\beta)$ and X is **discrete** with pmf $p(x|y,\alpha)$ $L(\alpha,\beta) = p(x^{(1)}|y^{(1)},\alpha) f(y^{(1)}|\beta) ... p(x^{(N)}|y^{(N)},\alpha) f(y^{(N)}|\beta)$

MLE

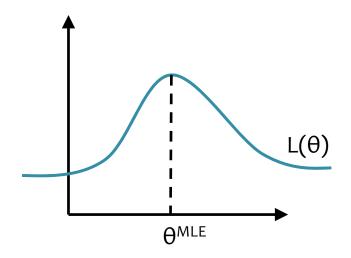
Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

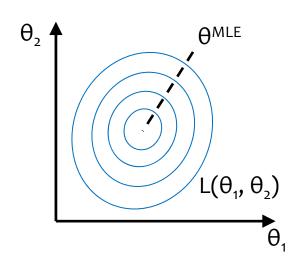
Principle of Maximum Likelihood Estimation:

Choose the parameters that maximize the likelihood of the data. N

$$\boldsymbol{\theta}^{\mathsf{MLE}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{i=1} p(\mathbf{x}^{(i)} | \boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)





MLE

What does maximizing likelihood accomplish?

- There is only a finite amount of probability mass (i.e. sum-to-one constraint)
- MLE tries to allocate as much probability mass as possible to the things we have observed...

... at the expense of the things we have not observed

Recipe for Closed-form MLE

Assume data was generated i.i.d. from some model 1. (i.e. write the generative story) $x^{(i)} \sim p(x|\theta)$

$$x^{(i)} \sim p(x|\theta)$$

Write log-likelihood 2.

$$\ell(\theta) = \log p(x^{(1)}|\theta) + \dots + \log p(x^{(N)}|\theta)$$

Compute partial derivatives (i.e. gradient) 3.

$$\frac{\partial \ell(\mathbf{\Theta})}{\partial \theta_{1}} = \dots$$

$$\frac{\partial \ell(\mathbf{\Theta})}{\partial \theta_{2}} = \dots$$

$$\frac{\partial \ell(\mathbf{\Theta})}{\partial \theta_{M}} = \dots$$

Set derivatives to zero and solve for $\boldsymbol{\theta}$ 4.

$$\theta^{\ell}(\theta)/\partial\theta_{m} = 0$$
 for all $m \in \{1, ..., M\}$
 θ^{MLE} solution to system of M equations and M variables

Compute the second derivative and check that $\ell(\theta)$ is concave down 5. at $\boldsymbol{\theta}^{\text{MLE}}$

MLE

Example: MLE of Exponential Distribution Goal:

- pdf of Exponential(λ): $f(x) = \lambda e^{-\lambda x}$
- Suppose $X_i \sim \operatorname{Exponential}(\lambda)$ for $1 \leq i \leq N$. Find MLE for data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

Steps:

- First write down log-likelihood of sample.
- Compute first derivative, set to zero, solve for λ .
- Compute second derivative and check that it is concave down at λ^{MLE} .

• pdf of Exponential(λ): f(x)• Suppose $X_i \sim \text{Exponential}(\lambda)$ for $1 \leq i \leq N$.
• Find MLE for data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

Example: MLE of Exponential Distribution

• First write down log-likelihood of sample.

$$\ell(\lambda) = \sum_{i=1}^{N} \log f(x^{(i)}) \tag{1}$$

$$= \sum_{i=1}^{N} \log(\lambda \exp(-\lambda x^{(i)}))$$
 (2)

$$=\sum_{i=1}^{N} \log(\lambda) - \lambda x^{(i)} \tag{3}$$

$$= N \log(\lambda) - \lambda \sum_{i=1}^{N} x^{(i)}$$
 (4)

MLE

- pdf of Exponential(λ): $f(x) = \lambda e^{-\lambda x}$
- Suppose $X_i \sim \mathsf{Exponential}(\lambda)$ for $1 \leq i \leq N$.
- Find MLE for data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

Example: MLE of Exponential Distribution

• Compute first derivative, set to zero, solve for λ .

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{d}{d\lambda} N \log(\lambda) - \lambda \sum_{i=1}^{N} x^{(i)}$$
 (1)

$$= \left(\frac{N}{\lambda}\right) - \left(\sum_{i=1}^{N} x^{(i)}\right) = 0 \tag{2}$$

$$\Rightarrow \lambda^{\mathsf{MLE}} = \frac{N}{\sum_{i=1}^{N} x^{(i)}} \tag{3}$$

MLE

In-Class Exercise

Show that the MLE of parameter ϕ for N samples drawn from Bernoulli(ϕ) is:

$$\phi_{MLE} = \frac{\text{Number of } x_i = 1}{N}$$

Steps to answer:

- Write log-likelihood of sample
- 2. Compute derivative w.r.t. φ
- 3. Set derivative to zero and solve for ϕ

MIF

Question:

Assume we have N samples $x^{(1)}$, $x^{(2)}, ..., x^{(N)}$ drawn from a Bernoulli(ϕ).

What is the **log-likelihood** of the data $\ell(\phi)$?

Assume
$$N_1 = \# \text{ of } (x^{(i)} = 1)$$

 $N_0 = \# \text{ of } (x^{(i)} = 0)$

Answer:

$$A. \quad I(\phi) = N_1 \log(\phi) + N_0 (1 - \log(\phi))$$

B.
$$I(\phi) = N_1 \log(\phi) + N_0 \log(1-\phi)$$

C. $I(\phi) = \log(\phi)^{N_1} + (1 - \log(\phi))^{N_1}$

C.
$$I(\phi) = \log(\phi)^{N_1} + (1 - \log(\phi))^{N_0}$$

D.
$$I(\phi) = \log(\phi)^{N_1} + \log(1-\phi)^{N_0}$$

E.
$$I(\phi) = N_0 \log(\phi) + N_1 (1 - \log(\phi))$$

F.
$$I(\phi) = N_0 \log(\phi) + N_1 \log(1-\phi)$$

G.
$$I(\phi) = \log(\phi)^{N_0} + (1 - \log(\phi))^{N_1}$$

H.
$$I(\phi) = \log(\phi)^{N_0} + \log(1-\phi)^{N_1}$$

1.
$$J(\phi) =$$
the most likely answer

MLE

Question:

Assume we have N samples $x^{(1)}$, $x^{(2)}$, ..., $x^{(N)}$ drawn from a Bernoulli(ϕ).

What is the **derivative** of the log-likelihood $\partial \ell(\theta)/\partial \theta$?

Assume
$$N_1 = \# \text{ of } (x^{(i)} = 1)$$

 $N_0 = \# \text{ of } (x^{(i)} = 0)$

Answer: A. $\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta} = \frac{\partial N_1}{\partial \theta} + (1 - \phi)N_0$ B. $\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta} = \frac{\phi}{N_1} + (1 - \phi)/N_0$ C. $\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta} = \frac{N_1}{\phi} + \frac{N_0}{N_0} / (1 - \phi)$ D. $\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta} = \frac{\log(\phi)}{N_1} + \frac{\log(1 - \phi)}{N_0}$ E. $\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta} = \frac{N_1}{\log(\phi)} + \frac{N_0}{\log(1 - \phi)}$

Learning from Data (Frequentist)

Whiteboard

Example: MLE of Bernoulli

MLE vs. MAP

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

Principle of Maximum Likelihood Estimation:

Choose the parameters that maximize the likelihood of the data. N

$$\boldsymbol{\theta}^{\mathsf{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1} p(\mathbf{x}^{(i)} | \boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)

Principle of Maximum a posteriori (MAP) Estimation:

Choose the parameters that maximize the posterior of the parameters given the data.

$$\boldsymbol{\theta}^{\mathsf{MAP}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^{N} p(\boldsymbol{\theta}|\mathbf{x}^{(i)})$$

Maximum a posteriori (MAP) estimate

MLE vs. MAP

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

Principle of Maximum Likelihood Estimation:

Choose the parameters that maximize the likelihood of the data. N

$$\boldsymbol{\theta}^{\mathsf{MLE}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{i=1} p(\mathbf{x}^{(i)} | \boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)

Principle of Maximum a posteriori (MAP) Estimation:

Choose the parameters that maximize the posterior of the parameters given the data.

Prior

Prior

$$\boldsymbol{\theta}^{\mathsf{MAP}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^{r} p(\mathbf{x}^{(i)}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

Maximum a posteriori (MAP) estimate

MLE vs. MAP

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$

Principle of Maximum Likeli

Choose the parameters that of the data.

$$\boldsymbol{\theta}^{\mathsf{MLE}} = \mathrm{arg}$$

Important!

Usually the parameters are continuous, so the prior is a probability density function

Maximum Likelihood Estimate (MLE)

Principle of Maximum a posteriori (MAP) Estimation:

Choose the parameters that maximize the posterior of the parameters given the data.

$$\boldsymbol{\theta}^{\mathsf{MAP}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

Maximum a posteriori (MAP) estimate

Learning from Data (Bayesian)

Whiteboard

- maximum a posteriori (MAP) estimation
- Example: MAP of Bernoulli—Beta

Recipe for Closed-form MLE

- 1. Assume data was generated i.i.d. from some model (i.e. write the generative story) $x^{(i)} \sim p(x|\theta)$
- 2. Write leg-likelihood (θ) $(\theta) = \log p(x^{(1)}|\theta) + ... + \log p(x^{(N)}|\theta)$
- 3. Compute partial derivatives (i.e. gradient)

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_{1}} = \dots$$
$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_{2}} = \dots$$
$$\dots$$
$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_{M}} = \dots$$

4. Set derivatives to zero and solve for $\boldsymbol{\Theta}$

$$\partial \ell(\theta)/\partial \theta_{\rm m} = {\rm o \ for \ all \ m} \in \{1, ..., M\}$$

 $\theta^{\rm MLE} = {\rm solution \ to \ system \ of \ M \ equations \ and \ M \ variables}$

5. Compute the second derivative and check that $\ell(\theta)$ is concave down at θ^{MLE}

Learning from Data (Bayesian)

Whiteboard

- maximum a posteriori (MAP) estimation
- Example: MAP of Bernoulli—Beta

Takeaways

- One view of what ML is trying to accomplish is function approximation
- The principle of maximum likelihood estimation provides an alternate view of learning
- Synthetic data can help debug ML algorithms
- Probability distributions can be used to model real data that occurs in the world (don't worry we'll make our distributions more interesting soon!)

Learning Objectives

MLE / MAP

You should be able to...

- Recall probability basics, including but not limited to: discrete and continuous random variables, probability mass functions, probability density functions, events vs. random variables, expectation and variance, joint probability distributions, marginal probabilities, conditional probabilities, independence, conditional independence
- 2. Describe common probability distributions such as the Beta, Dirichlet, Multinomial, Categorical, Gaussian, Exponential, etc.
- 3. State the principle of maximum likelihood estimation and explain what it tries to accomplish
- 4. State the principle of maximum a posteriori estimation and explain why we use it
- 5. Derive the MLE or MAP parameters of a simple model in closed form

NAÏVE BAYES

Naïve Bayes Outline

Real-world Dataset

- Economist vs. Onion articles
- Document → bag-of-words → binary feature vector

Naive Bayes: Model

- Generating synthetic "labeled documents"
- Definition of model
- Naive Bayes assumption
- Counting # of parameters with / without NB assumption

Naïve Bayes: Learning from Data

- Data likelihood
- MLE for Naive Bayes
- MAP for Naive Bayes
- Visualizing Gaussian Naive Bayes

Naïve Bayes

- Why are we talking about Naïve Bayes?
 - It's just another decision function that fits into our "big picture" recipe from last time
 - But it's our first example of a Bayesian Network and provides a clearer picture of probabilistic learning
 - Just like the other Bayes Nets we'll see, it admits
 a closed form solution for MLE and MAP
 - So learning is extremely efficient (just counting)

Fake News Detector

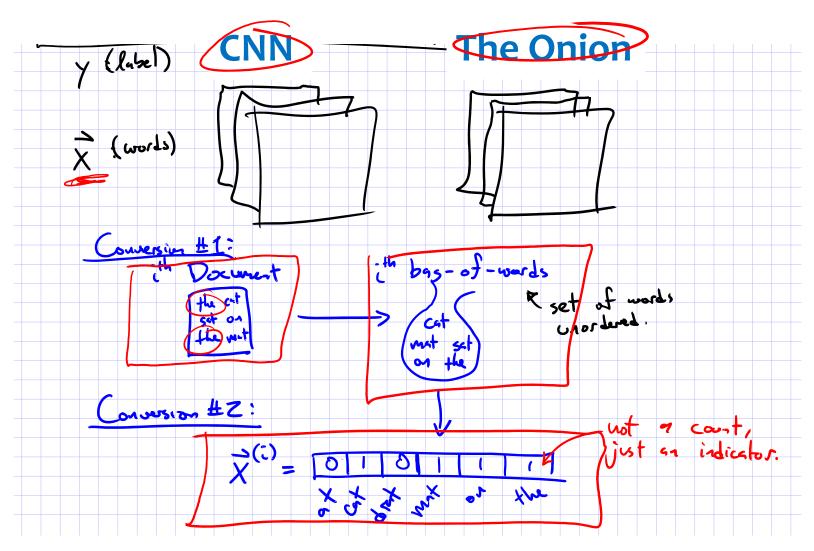
Today's Goal: To define a generative model of emails of two different classes (e.g. real vs. fake news)



The Onion



Fake News Detector



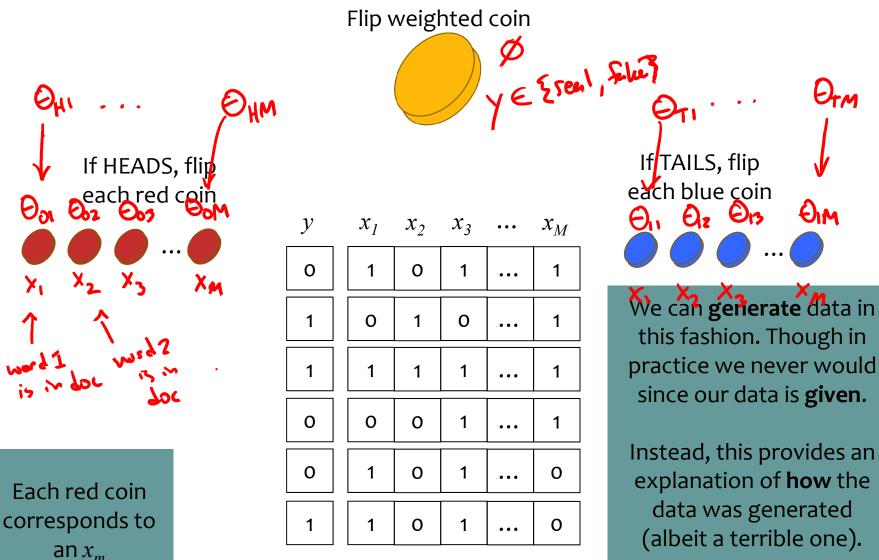
We can pretend the natural process generating these vectors is stochastic...

Naive Bayes: Model

Whiteboard

- Document → bag-of-words → binary feature vector
- Generating synthetic "labeled documents"
- Definition of model
- Naive Bayes assumption
- Counting # of parameters with / without NB assumption

Model 1: Bernoulli Naïve Bayes



practice we never would since our data is given.

DIM

Instead, this provides an explanation of how the data was generated (albeit a terrible one).

What's wrong with the Naïve Bayes Assumption?

The features might not be independent!!

- Example 1:
 - If a document contains the word "Donald", it's extremely likely to contain the word "Trump"
 - These are not independent!





NEWS IN BRIEF August 18, 2016 VOL 52 ISSUE 32 - Politics - Politicians - Election 2016 - Donald Trump

• Example 2:

If the petal width is very high,
 the petal length is also likely to
 be very high



Naïve Bayes: Learning from Data

Whiteboard

- Data likelihood
- MLE for Naive Bayes
- Example: MLE for Naïve Bayes with Two Features
- MAP for Naive Bayes

Recipe for Closed-form MLE

- 1. Assume data was generated i.i.d. from some model (i.e. write the generative story) $x^{(i)} \sim p(x|\theta)$
- 2. Write log-likelihood

$$\ell(\boldsymbol{\theta}) = \log p(x^{(1)}|\boldsymbol{\theta}) + \dots + \log p(x^{(N)}|\boldsymbol{\theta})$$

3. Compute partial derivatives (i.e. gradient)

$$\frac{\partial \ell(\mathbf{\Theta})}{\partial \theta_{1}} = \dots$$
$$\frac{\partial \ell(\mathbf{\Theta})}{\partial \theta_{2}} = \dots$$
$$\dots$$
$$\frac{\partial \ell(\mathbf{\Theta})}{\partial \theta_{M}} = \dots$$

4. Set derivatives to zero and solve for $\boldsymbol{\theta}$

$$\partial \ell(\Theta)/\partial \Theta_{\rm m} = 0$$
 for all m $\in \{1, ..., M\}$
 $\Theta^{\rm MLE} = {\rm solution \ to \ system \ of \ M \ equations \ and \ M \ variables}$

5. Compute the second derivative and check that $\ell(\theta)$ is concave down at θ^{MLE}