


Machine Learning Department
School of Computer Science
Carnegie Mellon University

# MLE/MAP

# +

# Naïve Bayes

Matt Gormley & Henry Chai
Lecture 16
Oct. 20, 2021

# Reminders

- Homework 5: Neural Networks
  - Out: Mon, Oct. 11
  - Due: Thu, Oct. 21 at 11:59pm
- Homework 6: Learning Theory / Generative Models
  - Out: Thu, Oct. 21
  - Due: Thu, Oct. 28 at 11:59pm
  - Same collaboration policy as Homework 3
    - Opt-in to homework groups on Piazza
  - IMPORTANT: you may only use 2 grace days on Homework 6
    - Last posible moment to submit HW6: Sat, Oct. 30 at 11:59pm

# (even more) Reminders

- **Midterm Exam 2**
  - Tue, Nov. 2, 6:30pm – 8:30pm
- **Practice for Exam 2**
  - Practice problems released on course website
    - (Tentatively) Out: Thu, Oct. 21
  - Mock Exam 2
    - (Tentatively) Out: Thu, Oct. 28
    - Due Sun, Oct. 31 at 11:59pm

# MLE AND MAP

# Likelihood Function

One R.V.

- Given N **independent, identically distributed (iid)** samples $D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$ from a **random variable** X …

- The **likelihood** function is
  - Case 1: X is **discrete** with probability mass function (*pmf*) $p(x|\theta)$
    $$L(\theta) = p(x^{(1)}|\theta)\, p(x^{(2)}|\theta) \ldots\ p(x^{(N)}|\theta)$$
  - Case 2: X is **continuous** with probability density function (pdf) $f(x|\theta)$
    $$L(\theta) = f(x^{(1)}|\theta)\, f(x^{(2)}|\theta) \ldots\ f(x^{(N)}|\theta)$$

The **likelihood** tells us how likely one sample is relative to another

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

# Likelihood Function

Two R.V.s

- Given N **iid** samples $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$ from a pair of **random variables** X, Y

- The **conditional likelihood** function:
  - Case 1: Y is **discrete** with *pmf* $p(y \mid x, \theta)$
    $$L(\theta) = p(y^{(1)} \mid x^{(1)}, \theta) \ldots p(y^{(N)} \mid x^{(N)}, \theta)$$
  - Case 2: Y is **continuous** with *pdf* $f(y \mid x, \theta)$
    $$L(\theta) = f(y^{(1)} \mid x^{(1)}, \theta) \ldots f(y^{(N)} \mid x^{(N)}, \theta)$$

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

# Likelihood Function

Two R.V.s

- Given N **iid** samples $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$ from a pair of **random variables** X, Y

- The **joint likelihood** function:
  - Case 1: X and Y are **discrete** with *pmf* $p(x,y|\theta)$
    $$L(\theta) = p(x^{(1)}, y^{(1)}|\theta) \ldots p(x^{(N)}, y^{(N)}|\theta)$$
  - Case 2: X and Y are **continuous** with *pdf* $f(x,y|\theta)$
    $$L(\theta) = f(x^{(1)}, y^{(1)}|\theta) \ldots f(x^{(N)}, y^{(N)}|\theta)$$
  - Case 3: Y is **discrete** with *pmf* $p(y|\beta)$ and X is **continuous** with *pdf* $f(x|y,\alpha)$
    $$L(\alpha, \beta) = f(x^{(1)}|\, y^{(1)}, \alpha)\, p(y^{(1)}|\beta) \ldots f(x^{(N)}|\, y^{(N)}, \alpha)\, p(y^{(N)}|\beta)$$
  - Case 4: Y is **continuous** with *pdf* $f(y|\beta)$ and X is **discrete** with *pmf* $p(x|y,\alpha)$
    $$L(\alpha, \beta) = p(x^{(1)}|\, y^{(1)}, \alpha)\, f(y^{(1)}|\beta) \ldots p(x^{(N)}|\, y^{(N)}, \alpha)\, f(y^{(N)}|\beta)$$

Mixed discrete/ continuous!

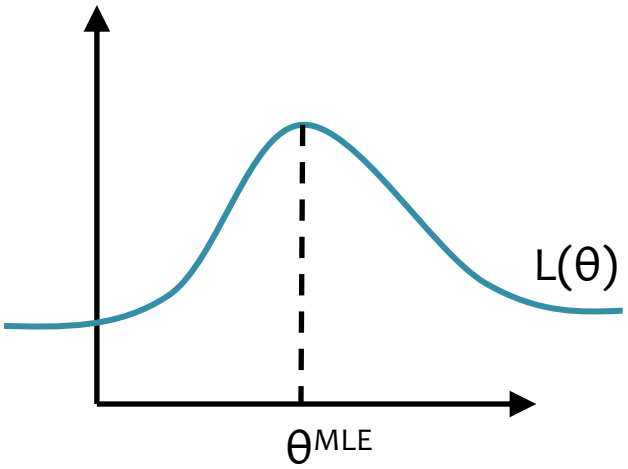
# MLE

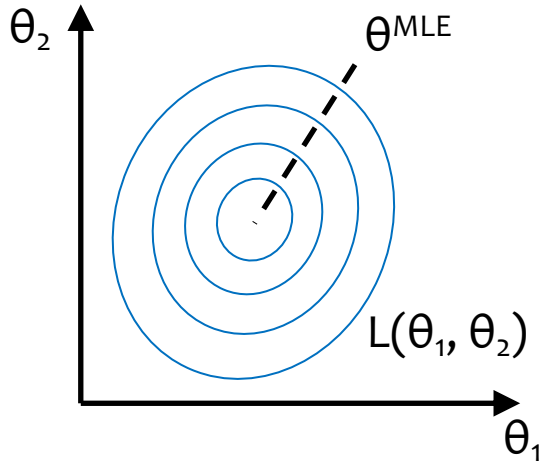
Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

**Principle of Maximum Likelihood Estimation:**
Choose the parameters that maximize the likelihood of the data.

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

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

1. Assume data was generated iid from some model, i.e., write the *generative story*

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

2. Write the log-likelihood

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

3. Compute partial derivatives, i.e., the gradient

   $\partial\ell(\boldsymbol{\theta})/\partial\theta_1 = \ldots$

   $\ldots$

   $\partial\ell(\boldsymbol{\theta})/\partial\theta_M = \ldots$

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

   $\partial\ell(\boldsymbol{\theta})/\partial\theta_m = 0$ for all $m \in \{1, \ldots, M\}$

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

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

# MLE of Exponential Distribution

*Whiteboard*

- Example: MLE of Exponential Distribution

# MLE

**In-Class Exercise**

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

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

**Steps to answer:**

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

# MLE

**Question:**

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

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

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

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

**Answer:**

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

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

C. $l(\phi) = \log(\phi)^{N1} + (1 - \log(\phi))^{No}$

D. $l(\phi) = \log(\phi)^{N1} + \log(1-\phi)^{No}$

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

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

G. $l(\phi) = \log(\phi)^{No} + (1 - \log(\phi))^{N1}$

H. $l(\phi) = \log(\phi)^{No} + \log(1-\phi)^{N1}$

I. $l(\phi) =$ the most likely answer

# Question 1

A

B

C

D

E

F

G

H

I

# MLE

**Question:**

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

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

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

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

**Answer:**

A. $\partial \ell(\boldsymbol{\theta})/\partial\theta = \phi^{N1} - (1 - \phi)^{No}$

B. $\partial \ell(\boldsymbol{\theta})/\partial\theta = \phi / N_1 - (1 - \phi) / N_o$

C. $\partial \ell(\boldsymbol{\theta})/\partial\theta = N_1 / \phi - N_o / (1 - \phi)$

D. $\partial \ell(\boldsymbol{\theta})/\partial\theta = \log(\phi) / N_1 - \log(1 - \phi) / N_o$

E. $\partial \ell(\boldsymbol{\theta})/\partial\theta = N_1 / \log(\phi) - N_o / \log(1 - \phi)$

F. $\partial \ell(\boldsymbol{\theta})/\partial\theta$ = *the derivative of the most likely answer*

# Question 2

A

B

C

D

E

F

# 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.

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

Maximum Likelihood Estimate (MLE)

**Principle of Maximum *a posteriori* (MAP) Estimation:**
Choose the parameters that maximize the posterior of the parameters given the data.

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

Maximum *a posteriori* (MAP) estimate

# MLE vs. MAP

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

**Principle of Maximum Likeli**

Choose the parameters that

of the data.

$$\boldsymbol{\theta}^{\text{MLE}} = \arg$$

Maximum Likelihood Estimate (MLE)

Important!

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

**Principle of Maximum *a posteriori* (MAP) Estimation:**

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

Prior

$$\boldsymbol{\theta}^{\text{MAP}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Maximum *a posteriori* (MAP) estimate

# Learning from Data (Bayesian)

*Whiteboard*

– *maximum a posteriori* (MAP) estimation

# Recipe for Closed-form MLE

1. Assume data was generated iid from some model, i.e., write the *generative story*

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

2. Write the log-likelihood

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

3. Compute partial derivatives, i.e., the gradient

   $\partial\ell(\boldsymbol{\theta})/\partial\theta_1 = \ldots$

   $\ldots$

   $\partial\ell(\boldsymbol{\theta})/\partial\theta_M = \ldots$

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

   $\partial\ell(\boldsymbol{\theta})/\partial\theta_m = 0$ for all $m \in \{1, \ldots, M\}$

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

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

# Recipe for Closed-form MAP

1. Assume data was generated iid from some model, i.e., write the *generative story*

   $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ and then for all i: $x^{(i)} \sim p(x|\boldsymbol{\theta})$

2. Write the log posterior

   $\ell_{MAP}(\boldsymbol{\theta}) = \log p(\boldsymbol{\theta}) + \log p(x^{(1)}|\boldsymbol{\theta}) + \ldots + \log p(x^{(N)}|\boldsymbol{\theta})$

3. Compute partial derivatives, i.e., the gradient

   $\partial \ell_{MAP}(\boldsymbol{\theta})/\partial \theta_1 = \ldots$

   $\ldots$

   $\partial \ell_{MAP}(\boldsymbol{\theta})/\partial \theta_M = \ldots$

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

   $\partial \ell_{MAP}(\boldsymbol{\theta})/\partial \theta_m = 0$ for all $m \in \{1, \ldots, M\}$

   $\boldsymbol{\theta}^{MAP}$ = solution to system of M equations and M variables

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

# Learning from Data (Bayesian)

*Whiteboard*

- 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…*

1. 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

- 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)

# Misinformation Detector

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

## Associated Press

### Steelers steady themselves behind linebacker T.J. Watt

By WILL GRAVES October 18, 2021

PITTSBURGH (AP) — Pittsburgh Steelers linebacker Devin Bush scooped up the loose ball and amid the chaos, immediately started running in the wrong direction before finding his bearings.

How very fitting for a team that's spent its first six weeks trying to figure things out.

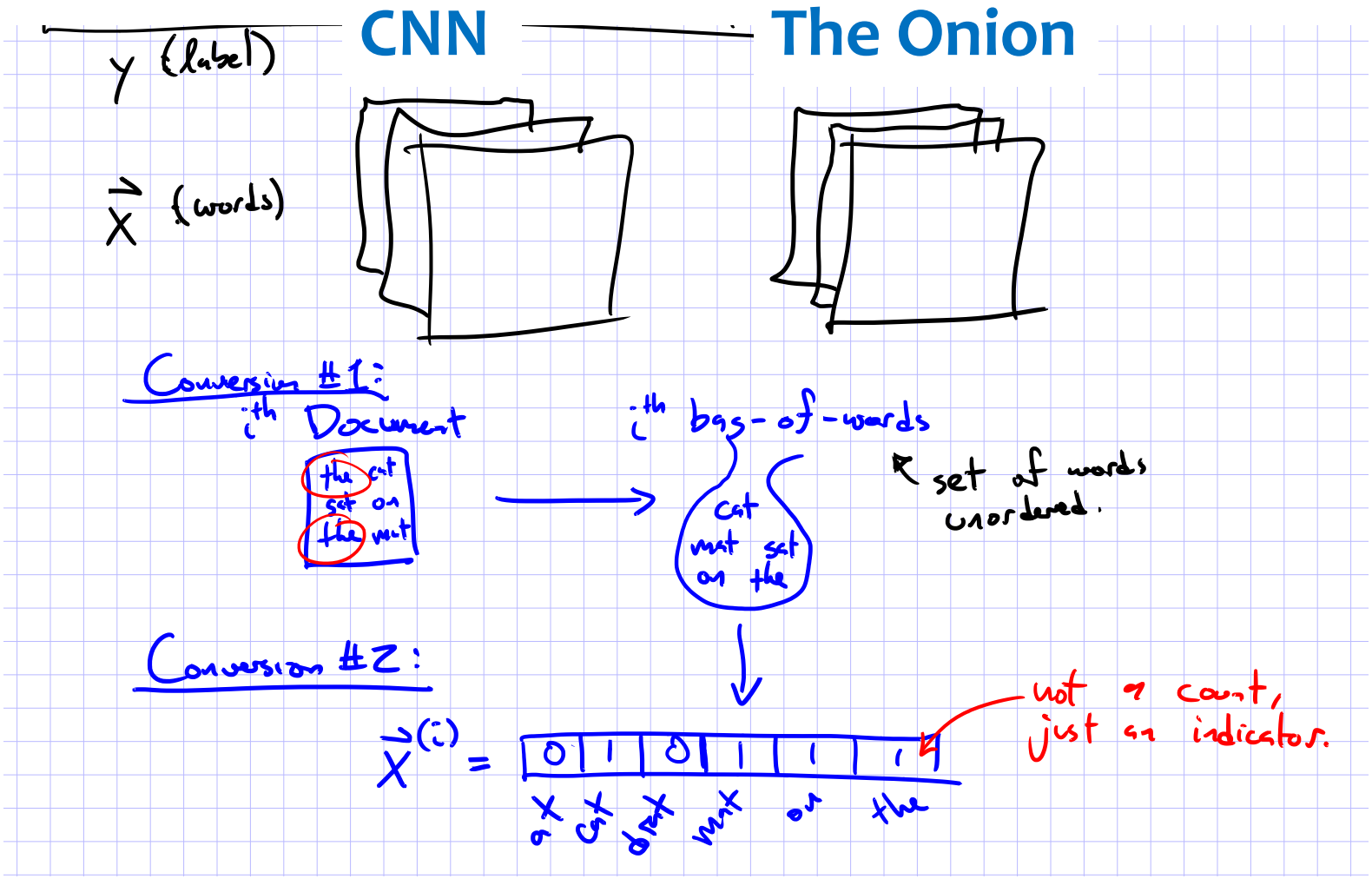
## The Onion

### Perfectly Preserved Fourth Watt Brother Discovered Frozen In Wisconsin Beer Cooler

Today 12:50PM | Alerts

WAUKESHA, WI—Hailing the massive specimen as the greatest NFL discovery of the century, league scientists announced Tuesday that they have discovered a perfectly preserved fourth Watt brother frozen in a Wisconsin beer cooler. "This is a historic find for football that could finally be the crucial missing link between J.J. and T.J.," said lead scientist Robin Grossman, who told reporters

# Fake News Detector




**We can pretend the natural process generating these vectors is stochastic…**

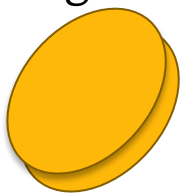
# Naive Bayes: Model

*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

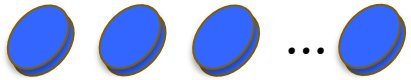
# Model 1: Bernoulli Naïve Bayes

Flip weighted coin



If HEADS, flip each red coin

If TAILS, flip each blue coin



| $y$ | $x_1$ | $x_2$ | $x_3$ | … | $x_M$ |
|---|---|---|---|---|---|
| 0 | 1 | 0 | 1 | … | 1 |
| 1 | 0 | 1 | 0 | … | 1 |
| 1 | 1 | 1 | 1 | … | 1 |
| 0 | 0 | 0 | 1 | … | 1 |
| 0 | 1 | 0 | 1 | … | 0 |
| 1 | 1 | 0 | 1 | … | 0 |

Each red coin corresponds to an $x_m$

We can **generate** data in this fashion. Though in practice we never would since our data is **given**.

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!


- 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 iid from some model, i.e., write the *generative story*

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

2. Write the log-likelihood

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

3. Compute partial derivatives, i.e., the gradient

   $\partial\ell(\boldsymbol{\theta})/\partial\theta_1 = \ldots$

   $\ldots$

   $\partial\ell(\boldsymbol{\theta})/\partial\theta_M = \ldots$

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

   $\partial\ell(\boldsymbol{\theta})/\partial\theta_m = 0$ for all $m \in \{1, \ldots, M\}$

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

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