Applied Machine Learning!!!

W207 Section 9
Rasika Bhalerao
rasikabh@berkeley.edu

Aug 23: Welcome!
Nov 8 and 22: No classes

Schedule

Supervised learning methods

	Sync	Topic
2	Aug 30	Linear Regression / Gradient Descent
3	Sep 6	Feature Engineering
4	Sep 13	Logistic Regression
5	Sep 20	Multiclass classification / Eval Metrics
6	Sep 27	Neural Networks
7	Oct 4	KNN, Decision Trees, Ensembles

Unsupervised learning methods

	Sync	Topic
8	Oct 11	KMeans and PCA
9	Oct 18	Text Embeddings
10	Oct 25	CNNs
11	Nov 1	EDA, Real data, Baselines
12	Nov 15	Fairness / Ethics
13	Nov 29	Fancy Neural Networks
14	Dec 6	Final Presentations

Assignment Schedule

Due Date	Assignment
Aug 28	HW1
Sep 4	HW2
Sep 11	HW3
Sep 18	HW4
Sep 25	HW5
Oct 2	HW6
Oct 16	Group project baseline
Oct 23	HW8
Nov 6	HW9
Nov 20	HW10
Dec 4	Final project notebook + presentation

Don’t forget to sign up for final project groups by week 4!

https://docs.google.com/document/d/1R3J_X1Rz6WP8eMQ2cyMC0wAr5iQdhMK_httdoNO6L0w/edit?usp=sharing

Behavior expectations

- Healthy disagreement is expected
- Be mindful of one another’s schedules
- Be a good listener
- Have fun in a professional manner
- Share related real-world experience
- Ask questions when something is confusing
- Keep it 100 but be respectful
- Be open-minded to new ideas in the real world and when coding
- On time for group meetings

View options (at the top) → Annotate

Async Practice Quiz Questions (vote!)

With n features and m training examples, the shape of the gradient should be:	(m, n)	(n, 1)	(m, 1)
Linear regression requires numeric inputs.	True	False	
A categorical feature with 10 possible values produces a one-hot representation with 10 features.	True	False	
Z-Score scaling can significantly improve training convergence speed.	True	False	
MAE for car price prediction could be improved simply by clipping negative predictions to 0.	True	False	

Feature Engineering, but first… Naive Bayes!

Spam?

Carol Christ Chancellor <CALmessages@berkeley.edu>
to calmessages_communication

Berkeley
UNIVERSITY OF CALIFORNIA

Chancellor's Office

Dear Campus Community,

With COVID-19 cases surging and positivity rates on the rise, it's clear we're in for a challenging January. As we navigate the omicron wave it's important that we be especially flexible and patient with one another. In this spirit, we're writing to share some updates to our instructional plans for the spring semester.

After consultation with the UC Berkeley public health committee, input from students, staff and faculty, and much deliberation, **we have decided to begin the semester with a two-stage process, with most courses being offered fully remote for the first two weeks (Jan. 18-28) and then moving to fully in-person instruction in the third week of the semester on Jan. 31.**

Some courses such as lab sections, studio courses, fieldwork, clinical courses, and graduate seminars may be taught in-person Jan 18-28. For these in-person courses, instructors may require in-person attendance but must offer appropriate make-up arrangements for students who are unable to

From: info.[redacted]@gmail.com
Sent on: Wednesday, April 1, 2020 12:38:23 PM
To:
Subject: Matter of Urgency, From United States Department of the Treasury

United States Department of the Treasury
Treasury Building 1500 Pennsylvania Avenue,
NW Washington, D.C., U.S
Financial Stability Oversight Council
(202) 622-88000
(425) 440-8440
Fax: (202) 622-64100
Hours: Mon-Fri 8:00am - 5:00pm
Us-treasury.gov

Good Day,

We have been instructed by the United Nation and US President to release all unclaimed ATM card to the beneficiary to curtails the recession because of the outbreak of the Corona Virus.

All we need from you is the delivery fee of $50. Immediately the fee is received, we will instruct the US POSTAL SERVICES to deliver it to you without any more delay.

[redacted]
Council chairman

https://info.phishlabs.com/blog/covid-19-phishing-update-nigerian-prince-lures-evolve-with-crisi

Positive or negative reviews?

Roll over image to zoom in

UFO Detector - Internal magnetometer interfaced with microcontroller for 24 hour/7 days a week

Cyphis

★☆☆☆☆ **One Star is Too Much for This Product**

Reviewed in the United States on September 7, 2012

I don't know if this is a scam or if mine was broken, but it doesn't work and I am still getting abducted by UFO's on a regular basis.

Amazon Customer

★★★★★ **it works**

Reviewed in the United States on March 21, 2018

Verified Purchase

it works great

Marcie G.

★★★★☆ **Freedom!**

Reviewed in the United States on December 12, 2013

Knowing when "they" are in the area this allows me to take off my tin foil hat a little more frequently. I've even peeked out the foil drapes occasionally, but still not that often.

I gave it 4 stars as I believe it could alert you sooner. I was almost teleported during one of my peeking episodes.

https://www.amazon.com/UFO-Detector-magnetometer-interfaced-microcontroller/dp/B000FVUKKO/ref=cm_cr_srp_d_product_top?ie=UTF8

What is the subject of this article?

MEDLINE Article

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

?

MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

https://web.stanford.edu/~jurafsky/slp3/slides/7_NB.pdf

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

Choosing a class:

$P(c|d5) \propto 3/4 * (3/7)^3 * 1/14 * 1/14 \approx 0.0003$

$P(j|d5) \propto 1/4 * (2/9)^3 * 2/9 * 2/9 \approx 0.0001$

https://web.stanford.edu/~jurafsky/slp3/slides/7_NB.pdf

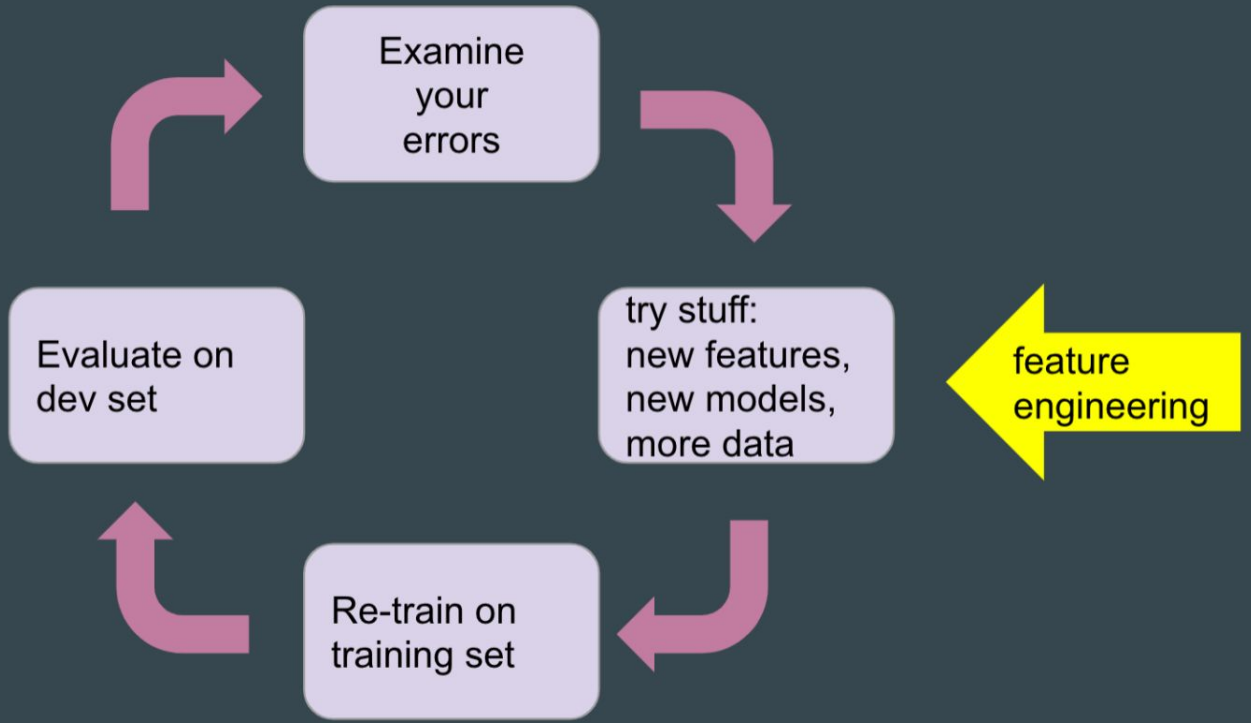
What if we are classifying images with pixels?

(Back to text) What would it look like if the features were independent given a class?

Suppose you get an email with the word “Squidward,” a word that is not in your training set. What happens now? And how do we fix this?

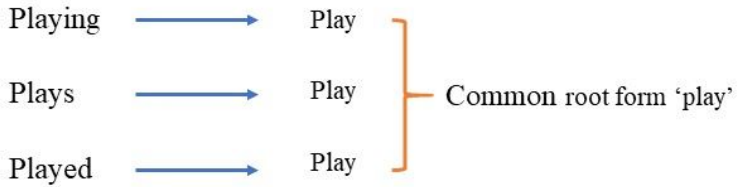
What is feature engineering / selection?

This is your life now



Stemming and Lemmatization

- Stemming
  - Keep a list of prefixes and suffixes, and cut them off words in the text
- Lemmatization
  - Use linguistics to get the root word
  - More complex but more useful



https://www.datacamp.com/community/tutorials/stemming-lemmatization-python

DataCamp

Feature engineering for text processing

- How you tokenize words
- Whether or not you ignore case
- How do you handle numbers? URLs? Timestamps? Usernames?

Why not just have as much detailed info as possible (keep case, all numbers, etc)?

View options (at the top) → Annotate

Old Async Practice Quiz Questions (vote!)

You can always assemble the joint distribution P(A,B) from the marginal distributions P(A) and P(B).	True	False
If A and B are conditionally independent given C, then A and B are independent.	True	False
If A and B are independent, then A and B are conditionally independent given C?	True	False
In our one-feature spam classifier, we have made no assumptions of independence.	True	False
Summing log probabilities is equivalent to multiplying probabilities.	True	False

View options (at the top) → Annotate

Practice Questions (not from your async)

What makes naive Bayes "naive"?	The assumption that the classes are independent.	**The assumption that the features are independent given the class.**	It usually doesn't work very well.
Feature selection is important because	**It keeps only the features that give the optimal performance.**	**It can remove poorly estimated features.**	
A perfectly calibrated classifier is ____% accurate on examples where the posterior probability is 85%.	**85**		

View options (at the top) → Annotate

Old Async Practice Quiz Questions

P(spam) = 0.4 P(“viagra”) = 0.05 P(“viagra” | spam) = 0.06

P(spam|“viagra”) =

(0.06 * 0.4) / (0.05) = 0.48

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

View options (at the top) → Annotate

Practice Questions

What is the Laplace (with k=1) smoothed estimate for P(sun) given this data:

domain: {sun,rain,wind}

observations: [sun,rain,rain,wind,sun,sun]

(3 + 1) / (6 + 3) = 4/9

Notebook!

To access later:
https://github.com/MIDS-W207/rasikabh/blob/main/live_sessions/Week3.ipynb