

Validating, Feature Engineering, and Other Improvements

David Quigley

CSCI 5622

2021 Fall

COMPUTING AND SOFTWARE

Career & Internship Fair

Tuesday, Oct. 5

11 am - 4 pm

VIRTUAL: On Handshake

**COMPLETE YOUR
PROFILE THEN SIGN UP!**

Employers from a variety of companies will be recruiting for computer science, software development or engineering and all levels. Speak with engineers one-on-one and learn about a broad spectrum of work opportunities.

colorado.edu/career/fairs-events

Accenture
Alarm.com
Altia
Amazon
AMERGINT Technologies
Big Compass
CACI
Caterpillar Inc.
Cigna
Cloud Campaign
Comcast - Central Division
CommScope
Congruex
CoreLogic
Credit One Bank
Danaher Corporation
DISH
Encompass Technologies
Epic
Expedia Group
Faegre Drinker Biddle & Reath LLP
FIS
Garmin
Hill AFB Civilian Engineering
ICR, Inc.
KPMG LLP
L3Harris Technologies

Lennox International
Lucid Software
Manifold
MiTek Inc.
National Security Agency (NSA)
Nelnet
NetApp
Omitron, Inc.
Palski & Associates, Inc.
Progressive Insurance
Qualcomm
Raytheon Technologies
Real-Time Innovations (RTI)
realtor.com
Ricoh USA, Inc
RSM US LLP
Sandia National Laboratories
Seagate Technology
Silicon Labs
Spectrum
Splunk
TASC
Trimble, Inc.
Visa, Inc.
Workiva
Xilinx

A graphic for a STEM career fair. It features a dark grey background with various white and yellow geometric shapes and icons. In the top left, there's a grid with mathematical symbols: a plus sign, a minus sign, a multiplication sign, and an equals sign. To the right of this is a yellow lightning bolt. Further right is a white rocket ship. Below the rocket is a yellow location pin. In the bottom left, there's a white Erlenmeyer flask with a yellow stopper. To the right of the flask is a yellow wrench. The word "STEM" is written in large, bold, white capital letters. Below it, the text "Industries Career & Internship Fair" is written in a smaller, white, sans-serif font. On the right side, there's a white box containing the date and time "October 7 11am - 4pm" and the text "Virtual: On Handshake Complete your profile then sign up!". Below this box, there's a paragraph of text about employer recruitment. At the bottom right of the box is a URL. The overall design is modern and tech-oriented.

STEM

Industries Career & Internship Fair

October 7
11am - 4pm

Virtual: On Handshake
Complete your
profile then sign up!

Employers will be recruiting for
computer science, robotics,
math, statistics, technology and
all engineering fields. CU Boulder
students and alumni from all
majors, experience levels and
backgrounds are welcome to
attend this FREE event.

colorado.edu/career/fairs-events



Career Services
UNIVERSITY OF COLORADO BOULDER



colorado.edu/career

Selected Employers Seeking Computing Talent:

Blue Horseshoe Solutions, Inc.
Capitalize
Colorado Dept of Transportation
Comcast - Central Division
CONMED
Credit One Bank
Deloitte
Esri
Fast Enterprises, LLC
Hill AFB Civilian Engineering
Hitachi ABB Power Grids
Holland & Hart LLP
Idaho National Laboratory
Jacobs
Johns Hopkins University Applied Physics
Keck Graduate Institute
Keysight Technologies
L3Harris Technologies
Lockheed Martin

Lumen
MIT Lincoln Laboratory
National Security Agency
(NSA)
NetApp
Parsons
Procter & Gamble (P&G)
Raytheon Technologies
RSM US LLP
Sandia National Laboratories
Seagate Technology
Silicon Labs
SK hynix NewCo -- New
Storage Solutions Company
Space Dynamics Laboratory
Spectrum
VMware, Inc.
Xcel Energy

Currently 79 total employers, see the full list on Handshake

Course Logistics

- Project Phase 2: Due 9/30 (Thursday!)
- Project Phase 2.1: Due 2(ish) weeks after Phase 2
 - i.e. Approximately 10/14 (we will distribute them manually and it will be due two weeks later).
- Problem Sets: ~~Problem Set 1 Feedback expected Thursday 9/23~~
 - ~~Currently anticipating a minor delay.~~
- Problem Set 2: Due 10/7

Today
11:50 AM

Who Will I Review For Phase 2.1?

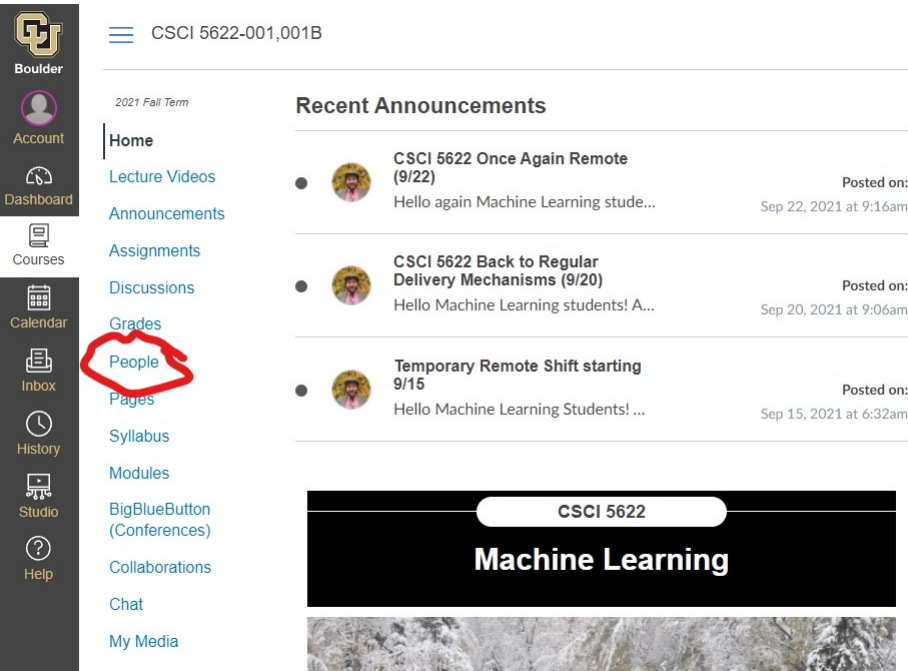
- <Your Project Group> + 1
- Except Group 26 will be reviewing Group 0 and Group 27 will be reviewing Group 26 (it's complicated, I'll email you)

What's <Your Project Group> For Me?

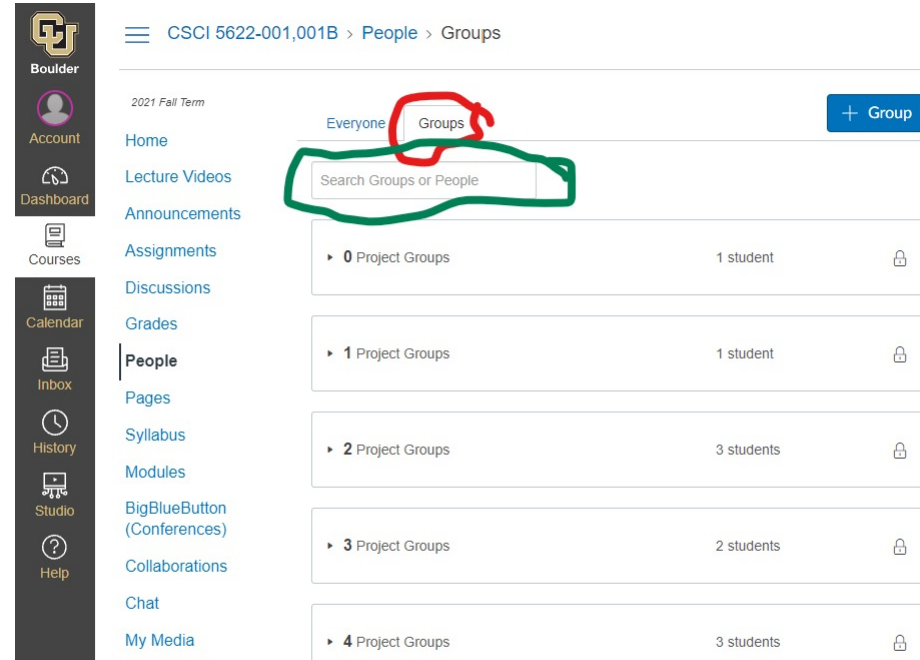


Viewing Project Groups

- Canvas → People (Blue Link on Left) → Groups (Tab on Top) → Search (Yourself)



The screenshot shows the Canvas interface for a course titled "CSCI 5622-001,001B". The left sidebar contains a list of navigation links: Home, Lecture Videos, Announcements, Assignments, Discussions, Grades, **People** (highlighted with a red circle), Pages, Syllabus, Modules, BigBlueButton (Conferences), Collaborations, Chat, and My Media. The main content area displays "Recent Announcements" for the 2021 Fall Term, including announcements about remote learning and a temporary shift starting on 9/15. At the bottom, there is a banner for "CSCI 5622 Machine Learning" with a background image of a snowy mountain.



The screenshot shows the Canvas interface for the same course, but with the "People" link selected in the sidebar. The top navigation bar shows "CSCI 5622-001,001B > People > Groups". The "Groups" tab is selected, and the search bar "Search Groups or People" is highlighted with a green circle. Below the search bar, a list of project groups is displayed, showing the number of groups, the number of students, and a lock icon.

Project Groups	Students	Lock Icon
0 Project Groups	1 student	🔒
1 Project Groups	1 student	🔒
2 Project Groups	3 students	🔒
3 Project Groups	2 students	🔒
4 Project Groups	3 students	🔒

But Also, Don't Worry!

- In Class:

- We watch everyone's pitch videos
- We ask questions about them
- Hopefully you're there to discover what kinds of questions get asked about your projects – it will help improve your project
- Hopefully you're there to get a head start on 2.1

- After Class:

- We (the Instructional Team) will distribute videos and papers from authors to reviewers.
- You will have two weeks (upon receiving digital copies of the materials, even if it isn't until Friday or the weekend) to review these materials and provide feedback.

Resources for Your Projects

- Some Groups have been asking about computing power above and beyond your laptops – we have that available!
 - Campus-wide ~~Systems Engineering~~ – OIT
 - Should be self-serve for basic needs
 - CS-Specific OpenStack Deployment
 - Email Me, I just put in a request for an instance for this course we should be able to share

Where We Left Off: Feature Engineering

Sequential Data

Sequential Data

I have a series of inputs that I want to combine to do classification

- We have a series of locations, can we gauge trajectory?
- We have 100 readings from an accelerometer, can we get gait?
- We have recent weather readings, can we better predict it?

Sequential Data

I have a series of inputs that I want to combine to do classification

- We have a series of locations, can we gauge trajectory?
- We have 100 readings from an accelerometer, can we get gait?
- We have recent weather readings, can we better predict it?

For those who need advanced techniques quickly for their projects, consider one of the textbooks for the class, Bayesian Reasoning & Machine Learning Ch. 23 - 26

Sequential Data

END

Timestamp	User	Action
100	David	Copy
150	David	Paste
200	David	Copy
250	David	Paste

Sequential Data – Past Actions as Features

Timestamp	User	PrevAction	<i>Action</i>
100	David	Login	<i>Copy</i>
150	David	Copy	<i>Paste</i>
200	David	Paste	<i>Copy</i>
250	David	Copy	<i>Paste</i>

Sequential Data – Past Actions as Features

Timestamp	User	PrevAction	<i>Action</i>
100	David	Login	<i>Copy</i>
150	David	Copy	<i>Paste</i>
200	David	Paste	<i>Copy</i>
250	David	Copy	<i>Paste</i>

Can this scale? Does this give us enough info?

Sequential Data – Past Actions as Features

Timestamp	User	N-2Action	PrevAction	<i>Action</i>
100	David	N/A	Login	<i>Copy</i>
150	David	Login	Copy	<i>Paste</i>
200	David	Copy	Paste	<i>Copy</i>
250	David	Paste	Copy	<i>Paste</i>

You could do this forever...

Curse of Dimensionality

+ samples

Features have a *value* and a *cost*

- Value: benefit to your training / accuracy
- Cost: Computation time, memory
Time & Effort to collect

If you're going to include a feature, it better be worth it!

Lots of ways to evaluate / measure the *usefulness* of a feature...

<https://www.kdnuggets.com/2017/04/must-know-curse-dimensionality.html>

Curse of Dimensionality – Leave-one-out

Train & evaluate your classifier while leaving out one feature at a time.

Log Sq Ft	# Bedrooms	House Lat.	House Long.	Cost
3.2	4	121.33	47.34	500K
3.09	3	121.33	55.23	450K
2.87	2	121.33	55.34	200K
3.06	2	130.99	47.34	1500K

~~Acc: 58%~~
777

Curse of Dimensionality – Leave-one-out

Train & evaluate your classifier while leaving out one feature at a time.

Log Sq Ft	# Bedrooms	House Lat.	House Long.	Cost
3.2	4	121.33	47.34	500K
3.09	3	121.33	55.23	450K
2.87	2	121.33	55.34	200K
3.06	2	130.99	47.34	1500K

Acc: 68%

???

Curse of Dimensionality – Leave-one-out

Train & evaluate your classifier while leaving out one feature at a time.

Log Sq Ft	# Bedrooms	House Lat.	House Long.	Cost
3.2	4	121.33	47.34	500K
3.09	3	121.33	55.23	450K
2.87	2	121.33	55.34	200K
3.06	2	130.99	47.34	1500K

~~Acc: 65%~~

777

Curse of Dimensionality – Leave-one-out

Train & evaluate your classifier while leaving out one feature at a time.

Log Sq Ft	# Bedrooms	House Lat.	House Long.	Cost
3.2	4	121.33	47.34	500K
3.09	3	121.33	55.23	450K
2.87	2	121.33	55.34	200K
3.06	2	130.99	47.34	1500K

Acc: 78%
↑↑↑

Curse of Dimensionality – Feature Optimization

Select-K-Best – Leave out all but K features, with the K features being the highest scoring on your test

Selecting to a threshold – you can define a threshold of some sort to determine what features are “good enough”

https://scikit-learn.org/stable/modules/feature_selection.html

Curse of Dimensionality - Sparse Data

Rating-GoT	Rating-BB	Rating-F	Rating-AD
0	0	0	0
4	2	5	4
5	0	0	0
0	5	0	0

Curse of Dimensionality – One-hot Encoding

How many “colors” are there?

- in your crayon box?
- in computational space ($256 * 256 * 256$)
- for a dog?

What benefits do you get from granularity of color?

What problems do you get from granularity of color?

Categorical Data – Feature Hashing

- 1) Choose however many features you're willing to accept (n)
 - based on size constraints, etc.
- 2) Create a unique hash function to encode your categories
 - Each one should be unique

Now you have a set of n -dimensional vectors that represent the variable*

*concerns include collisions, interpretation from classifiers

Categorical Data – Feature Binning

Even if they don't have *order*, features may have *clusters*.

- Copy / Paste / Cut are “edit” actions
- Typing letters are “generate” actions

Put these items into clusters before you encode these features!

This should be motivated by some theory about your data / problem!

Curse of Dimensionality – Sequential Data

How far back do we want to keep our window?

Do we want to keep our expanding feature set?

Sequential Data - Categorization

- 1) Build a separate category for each possible action sequence (or action cluster, see binning)
 - 1) Edit, edit edit; edit, edit, generate; edit, generate, edit; ...
- 2) Build a separate feature for each kind of sequence.

Evaluating Models

Back to it!

Types of Errors (Week 2)

<i>Classified As</i>	C	~C
<i>Ground Truth</i>		
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

Types of Errors

Predicted Positive Rate (Precision) = Hits / (Hits + False Alarm)

<i>Classified As</i>	C	~C
<i>Ground Truth</i>		
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

Types of Errors

True Positive Rate (Sensitivity, Recall) = Hits / (Hits + Miss)

<i>Classified As</i>	C	~C
<i>Ground Truth</i>		
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

Types of Errors

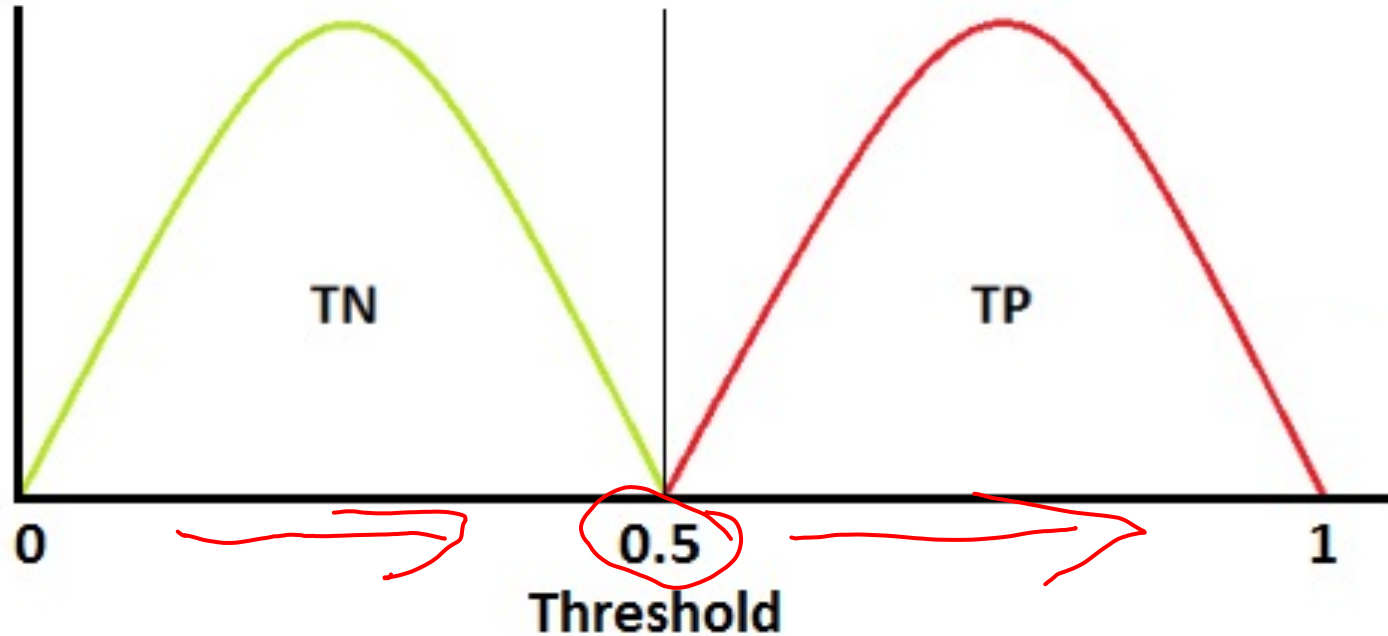
False Positive Rate (Specificity) = $\text{Corr. Rej.} / (\text{False Alarm} + \text{Corr. Rej.})$

<i>Classified As</i>	C	~C
<i>Ground Truth</i>		
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

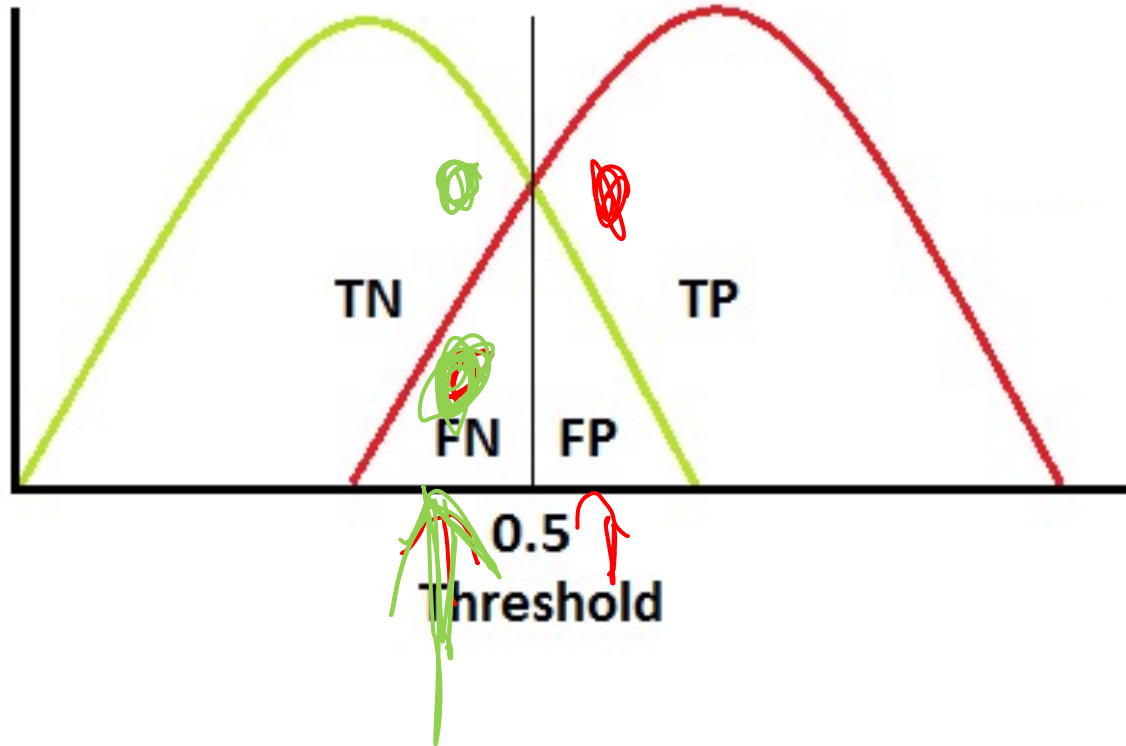
Confusion Matrix

<i>Classified As</i>	C	~C
<i>Ground Truth</i>		
C	Count of True Positives (Hit)	Count of False Negatives (Miss)
~C	Count of False Positives (False Alarm)	Count of True Negatives (Correct Rej.)

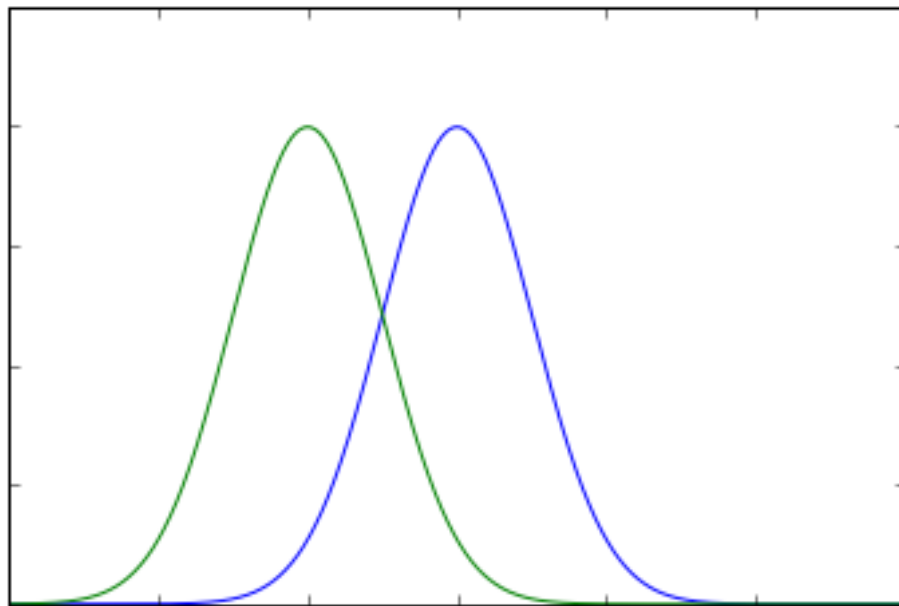
Classification Errors – Regression Model



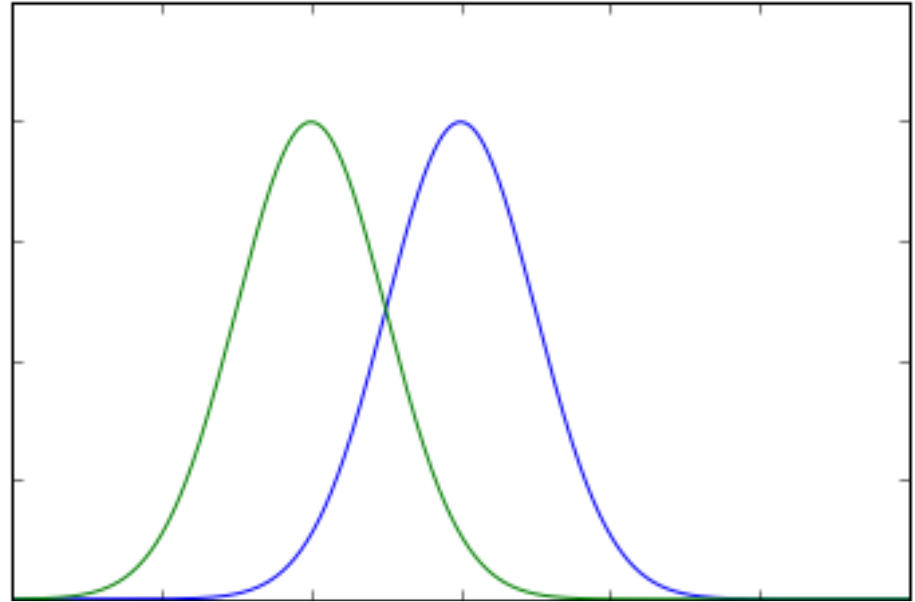
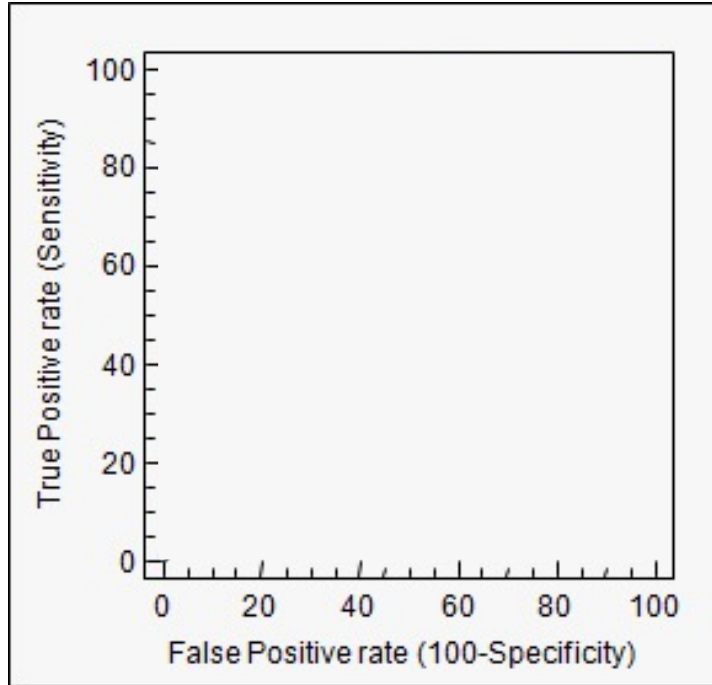
Classification Errors – Regression Model



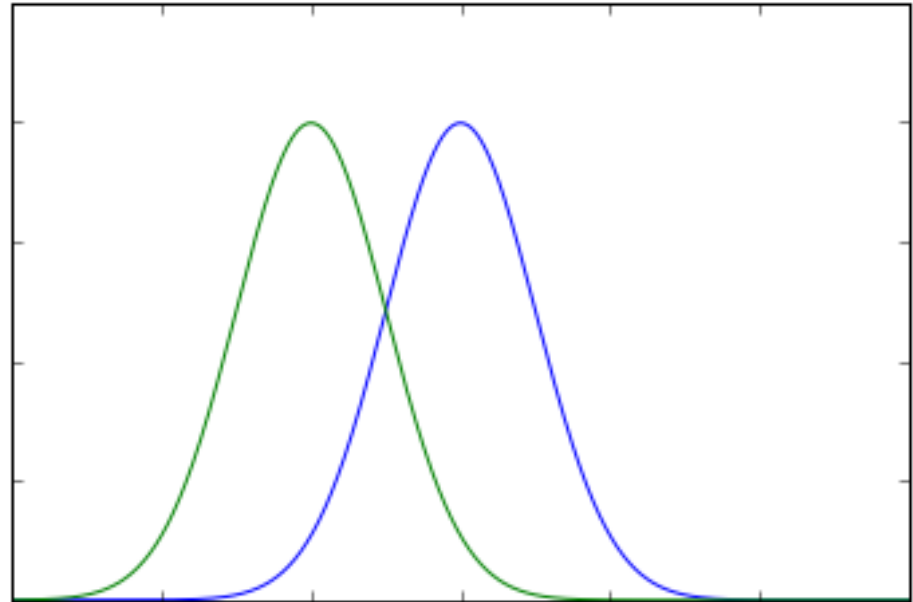
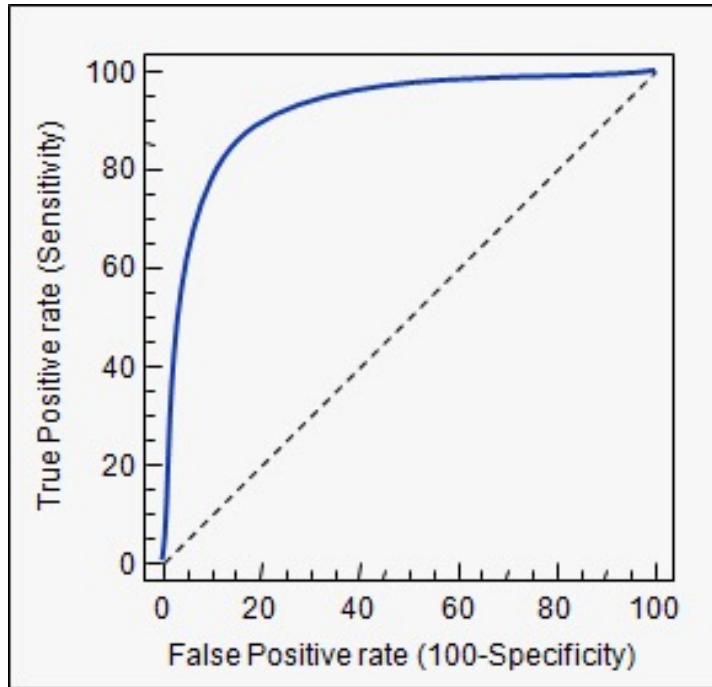
Gaussian Distributions



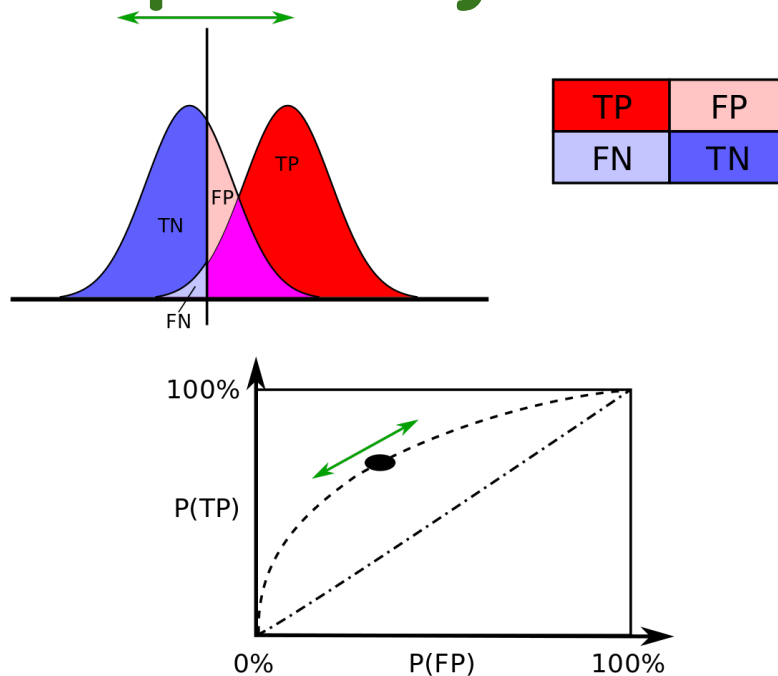
Classification Errors – Regression Model



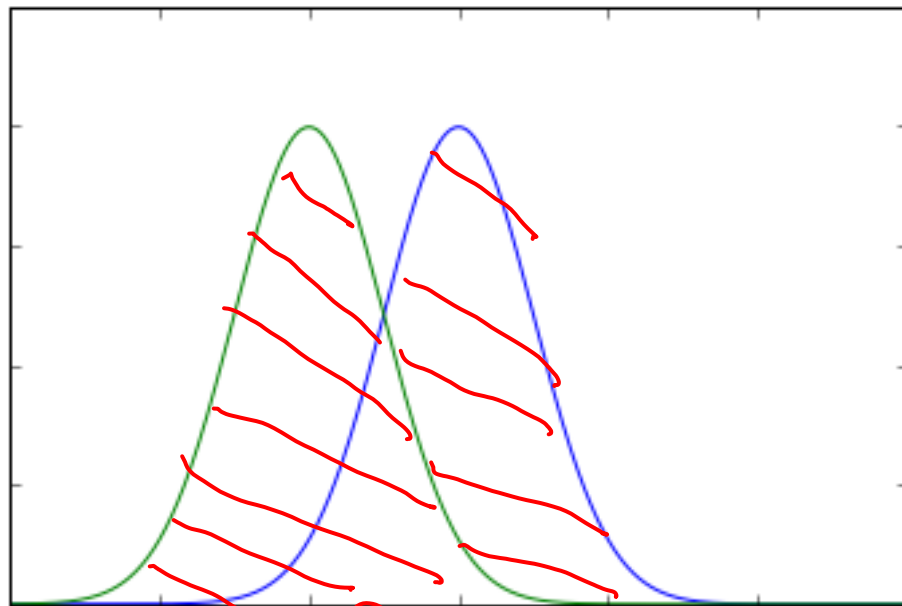
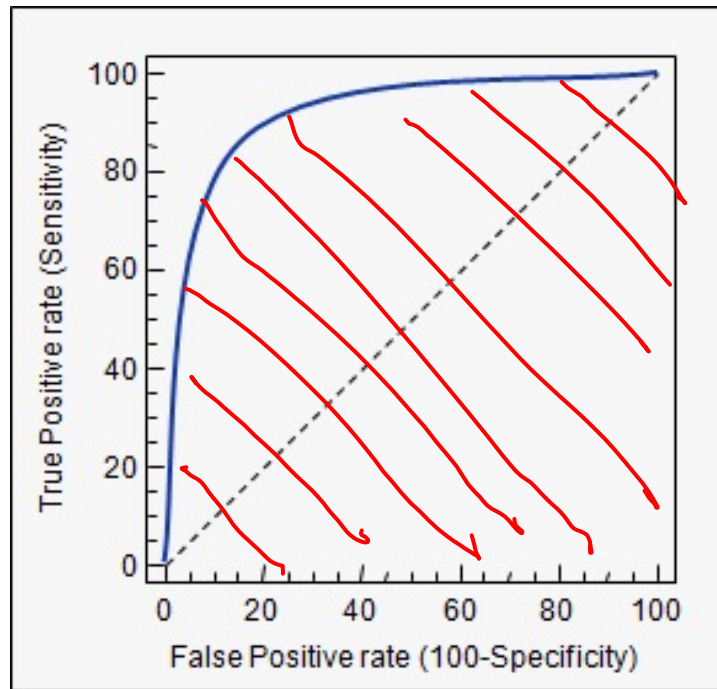
Classification Errors – ROC Curve (receiver operating characteristic)



Adapting True Pos rate vs. False Pos rate (sensitivity vs. specificity)

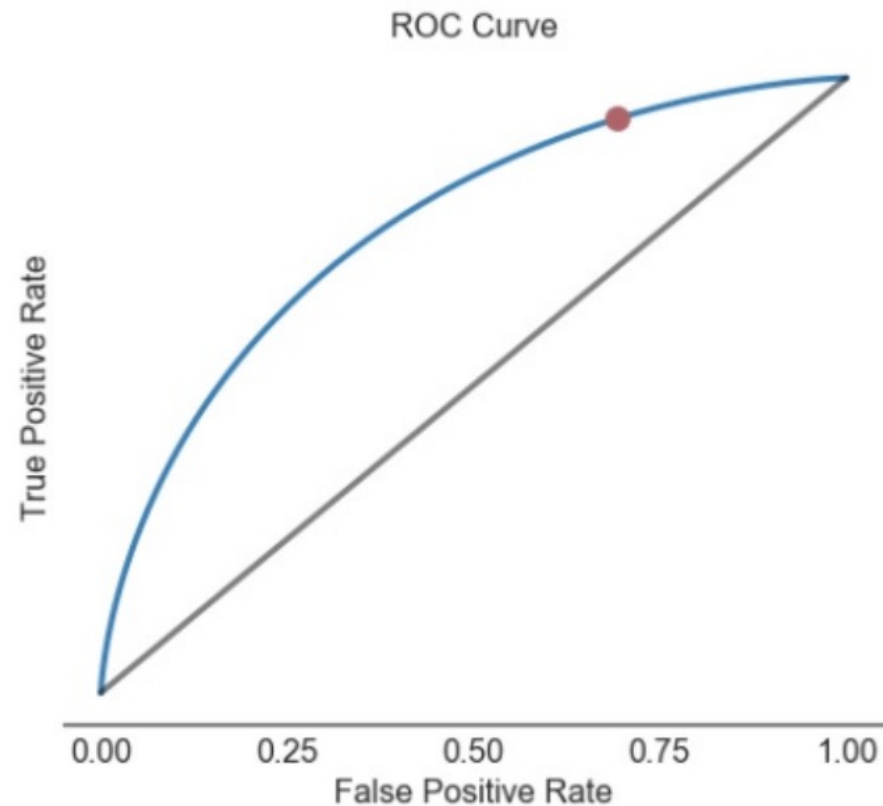
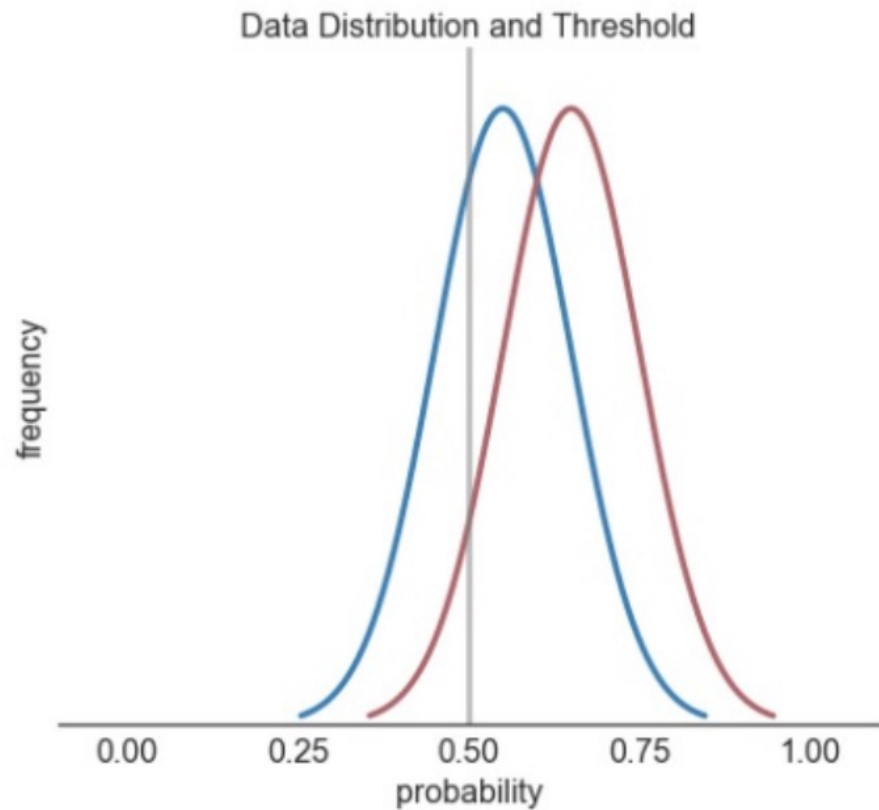


ROC Curve vs AUC

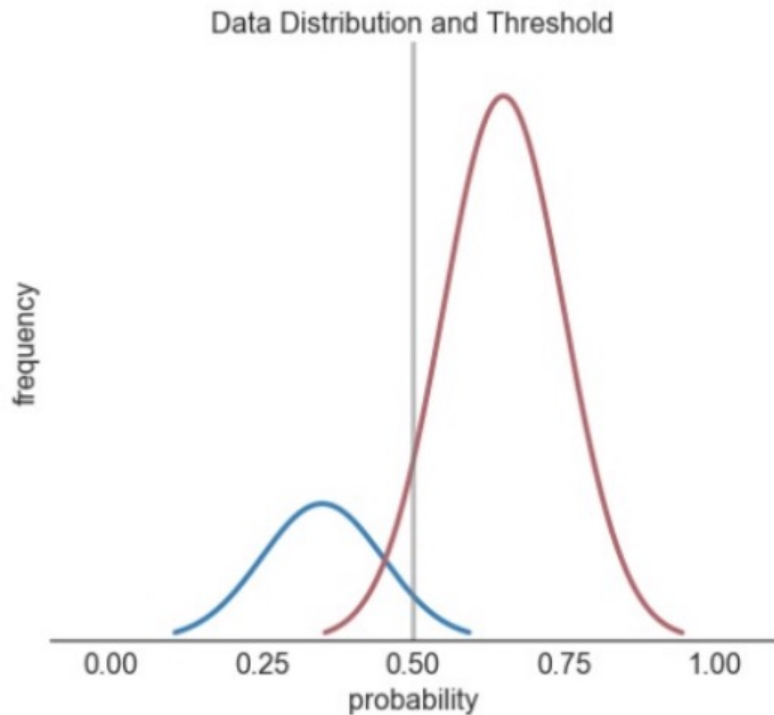


$n \sigma + d c$

Sharpness of ROC Curve



Non-Identical Distributions - Skew



Non-Identical Distributions - Skew

END

