Goodness of Fit!

Class 6: AISC 1003
Instructor: Bhavik Gandhi

Course Logistics

Good question!!

Course Plan*

Week	Date	Topic	Activity
1	May 13, 2023	Intro to AI, ML, DS	Group Exercise (Risk and Applications)
2	May 20, 2023	Types of ML	Practice Exercise,
3	May 27, 2023	Linear Regression - OLS	Group Activity (Excel OLS)
4	Jun 3, 2023	Linear Regression - Gradient Descent	Indiv Assignment 1
5	Jun 10, 2023	Regularization	Group Exercise (Model building)
6	Jun 17, 2023	Cross Validation and Goodness of Fit	Group Assignment 2
7	Jun 24, 2023	Mid-Term	Mid-Term

*Course Plan is tentative and subject to change. All assessments are in-class assessments. No extensions will be provided.

Course Plan*

Week	Date	Topic	Activity
9	Jul 8, 2023	Classification, Sigmoid Function	Group Exercise (GoF)
10	Jul 15, 2023	Logistic Regression	Group Assignment 3
11	Jul 22, 2023	SVM	Indiv Assignment 4, Group Exercise (SVM)
12	Jul 29, 2023	SVM Kernels	Group Assignment 5 (SVM)
13	Aug 5, 2023	Sensitivity Analysis, Model Interpretability	Group Assignment 6 (Sensitivity Analysis)
14	Aug 12, 2023	Final	Final
15	Aug 19, 2023	~~Decision Trees, Ensembling~~	Individual Exercise

*Course Plan is tentative and subject to change. All assessments are in-class assessments. No extensions will be provided.

Grading & Class Structure

Grades will be absolute

Any integrity violations will get you 0 on the assessment

Bonus points can help you improve your grade

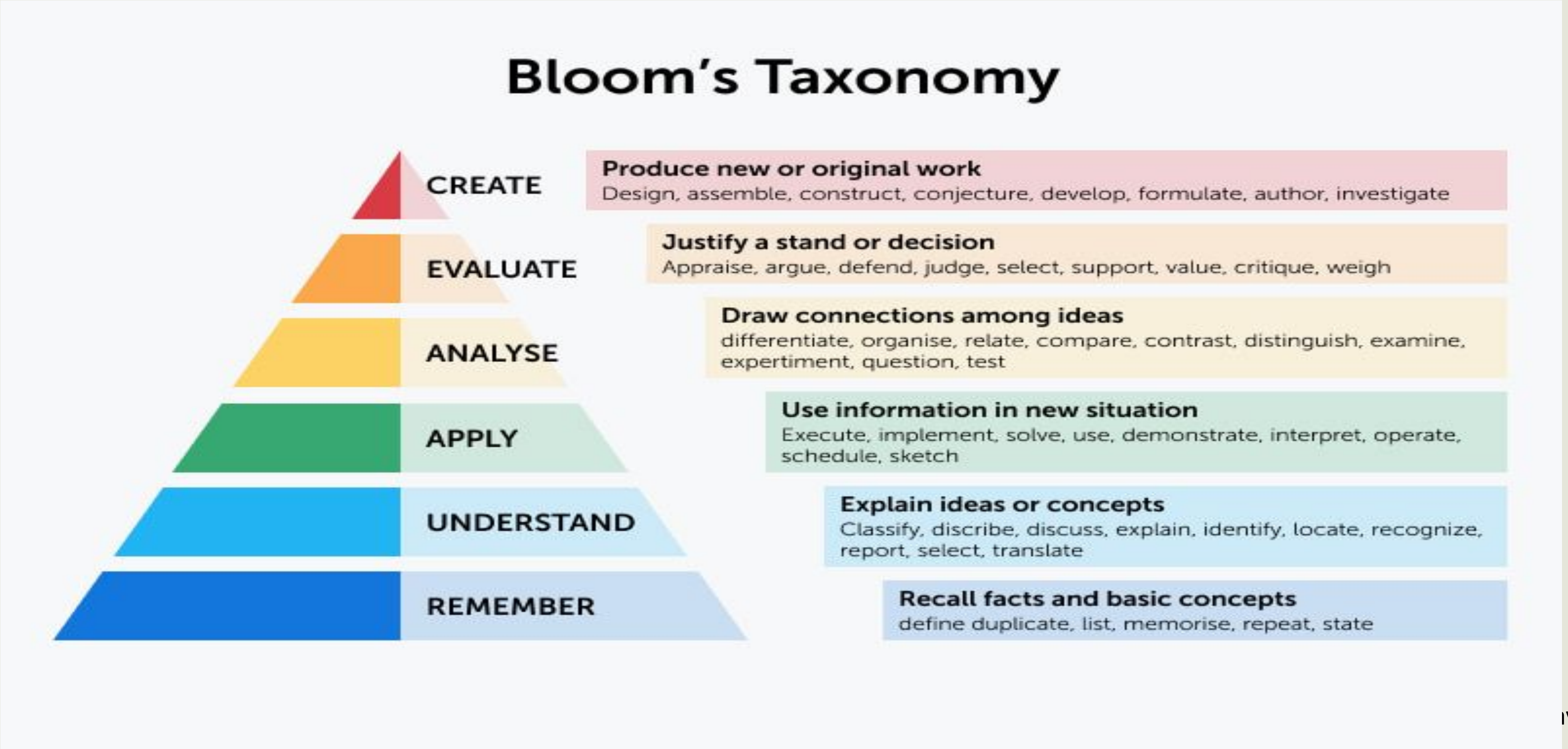
But you will still have to work! 😅

All submissions should be accompanied by an accurate task status report detailing which (and how well) the tasks/subtasks were completed and by who (in case of a group assignment) which can serve as a MECE table.

The class will usually be 1 - 1.5 hrs on instruction

The remaining 2-2.5 hrs will be on application and analysis of the material and in-class activities

Bloom’s Taxonomy



Slides, Learning Resources etc

Slides used here are only for reference, further asynchronous slides will be posted. However assessments will only be on the taught material (understanding, application, analysis, evaluation etc. - check Bloom's taxonomy)

Various concepts will be explained using whiteboard, so feel free to take notes

Additional resources will be posted in Moodle.

Rubrics may be used to evaluate assignments. The assignments will not have bonus pts but upto 10% bonus pts may be awarded for extraordinary work as per the instructor's discretion.

Class Decorum

No talking in the class even to yourself even if it is related to the subject matter except during breaks and group assignments/exercises.

Please step out for at least 15 minutes if you need to talk. If you are found talking you will be asked to step out and instruction to the class will be paused until you step out.

No language other than English in class even during breaks or group work. You need to get English and AI-ready. If found talking in another language you will be asked to leave the class until the next break and instruction to the class will be paused until you do.

No mobile phone ringing/vibrating in class. No recording in the class. If either of these happen you will be asked to leave the class until the next break and instruction will be paused until you do.

You can join in late in class, just come in and close the door. However attendance may be missed, you will miss on some content/instruction and you will not be allowed to join in late if an assessment has begun. Also, you will not get any make-up opportunities so be on time.

Email Writing

It is important to learn how to write good professional emails. Below are some useful instructions to bear in mind.

Please make sure any emails you write are to the point and clearly highlight what you expect from the person addressed.

Please make sure you do the groundwork and provide all the necessary details as opposed to expecting the person addressed to do the work.

Please make sure the subject of the email clearly identifies the objective of the email.

If it is an fyi email, please mention so right away. If you expect a certain action, please mention briefly and exactly what action do you expect and why is it objectively justified.

Academic Integrity

Academic Integrity is extremely important, honesty is highly valued in Canada

- Cannot accept any submissions over email
- If you face any issues submitting the assignment, follow the below steps
  - Take screenshots
  - Send an email to bhavik.gandhi@tbcollege.com with your complete assignment attached
  - Email itsupport@tbcollege.com or moodlesupport@tbcollege.com with these screenshots and get it resolved asap
  - Late submissions will incur a penalty so budget your time accordingly, genuine issues as verified by the it team or moodle team will be granted penalty-free extension
- Read the instructions and do not indulge in plagiarism. Anti-plagiarism softwares are very smart
- Please do not make me file ADRs 🙏 The risk is not worth the few extra pts you may get
- Not following instructions during test, even if you are not explicitly cheating, will still result in an ADR

Bonus Points!

- Total Bonus Points for the course will be capped at 20 (20%), total points capped at 100
- Attendance, Participation/Contribution: Upto 6 bonus points at the instructor’s discretion (includes participation in class, exercises, forums etc.). Will call it out in forums, exercise feedback or by email etc.
- Slide Corrections/Suggestions: Upto 4 bonus points at the instructor’s discretion, kindly email suggestions/corrections and save the response emails.
- Upto 10 bonus points distributed across assignments, exams and project.
- Bonus Point Assignment: Within 2 days of the last/W15 class each student will be required to submit how many bonus points they have and how they’d like their bonus points to be assigned. Failure to do so may lead to forfeiture of your bonus points. Set reminders now!

Group Exercises and Assignments!

- Points will be a combination of group performance and individual performance

- Individual responsibilities should be clearly outlined by the group in a MECE format in the assignment/exercise report

- The best submitted solutions at the instructor’s discretion may be shared with the class. You may get some bonus pts for knowledge sharing is your submission in a graded assessment is selected for sharing. Please write to me if you have any concerns regarding this.

Success Factors

No one can teach you all of Machine Learning

There's just too much to teach, so the idea of this course is to give you a primer into these fields so that you can explore them correctly.

To succeed in this course, apart from following instructions and having a strong sense of integrity, you will need

1. Basic Mathematical Skills
2. Problem Solving
3. Time Management
4. Group Working Skills
5. Practice

Today's Outline!

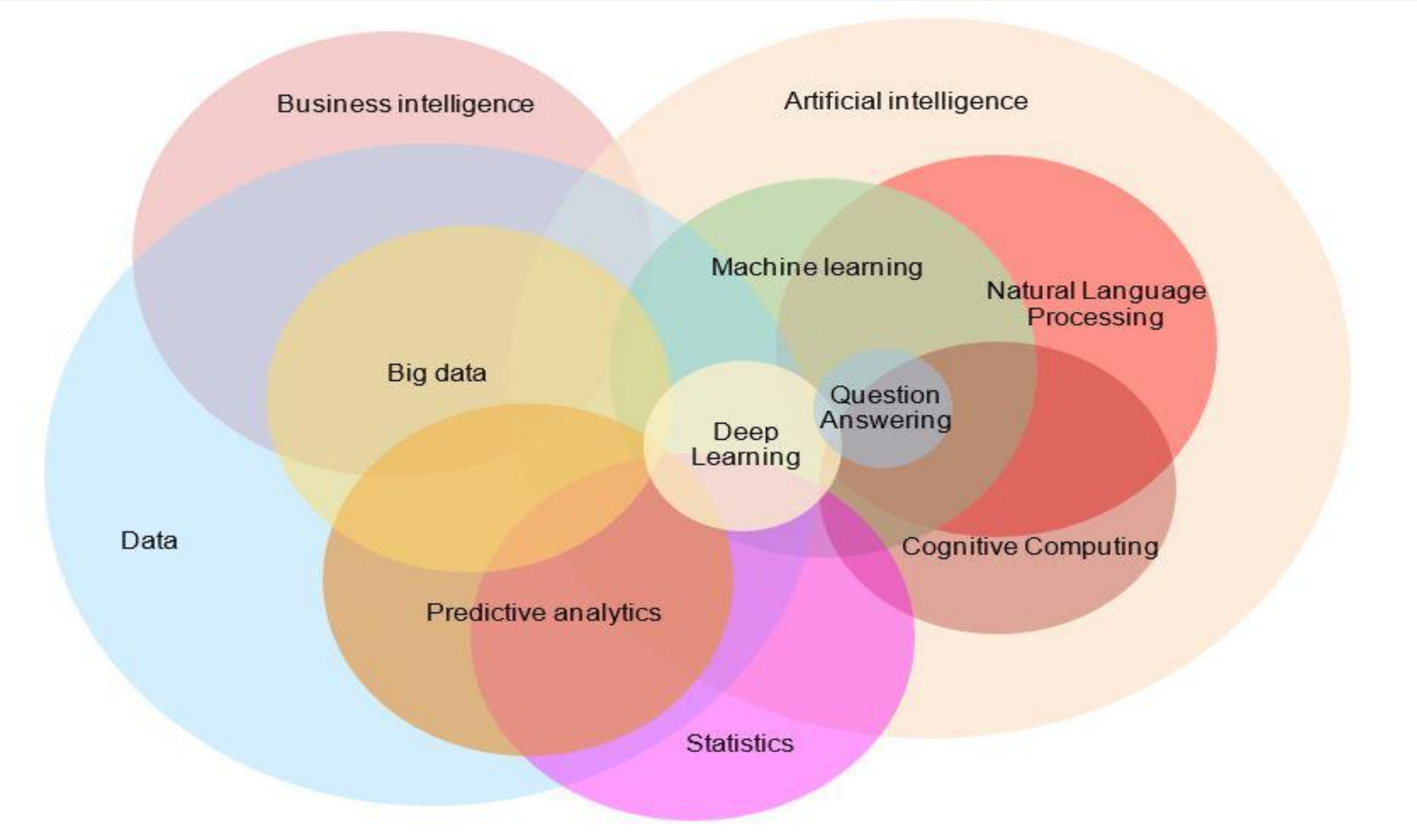
- k-fold Cross-Validation
- ROC Curves
- F-measures
- Group Assignment

Review

What did we talk about last time?

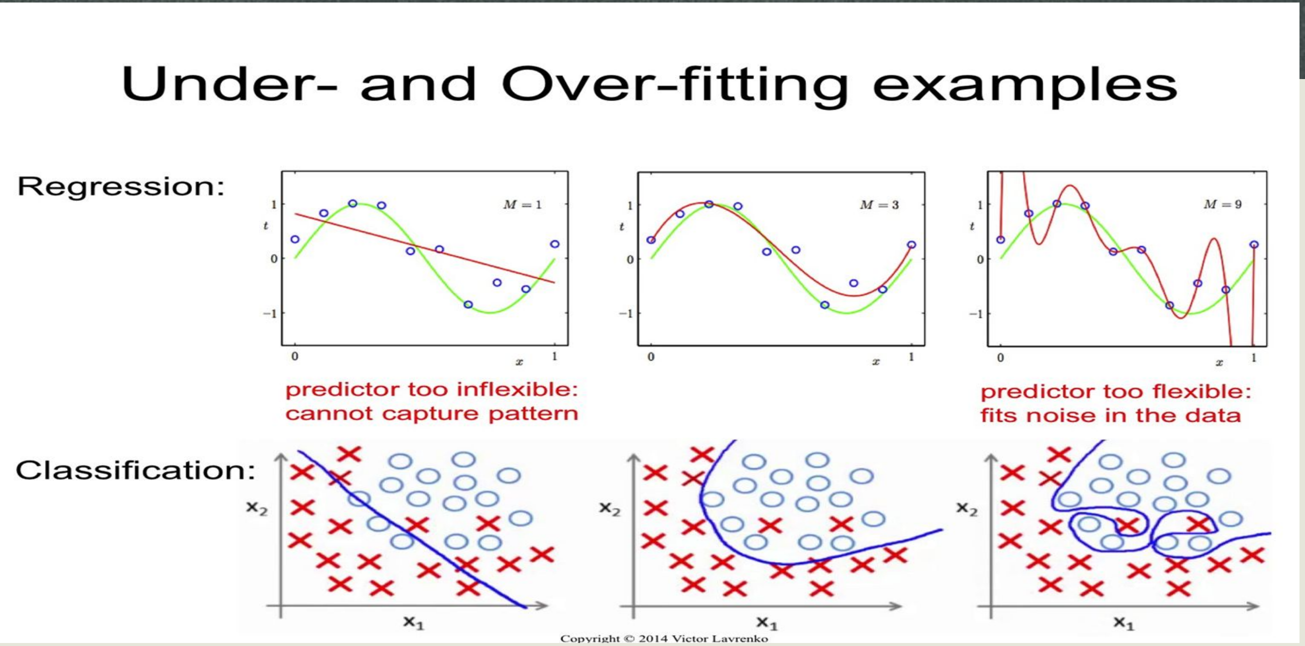
Intelligence, Learning & Challenges

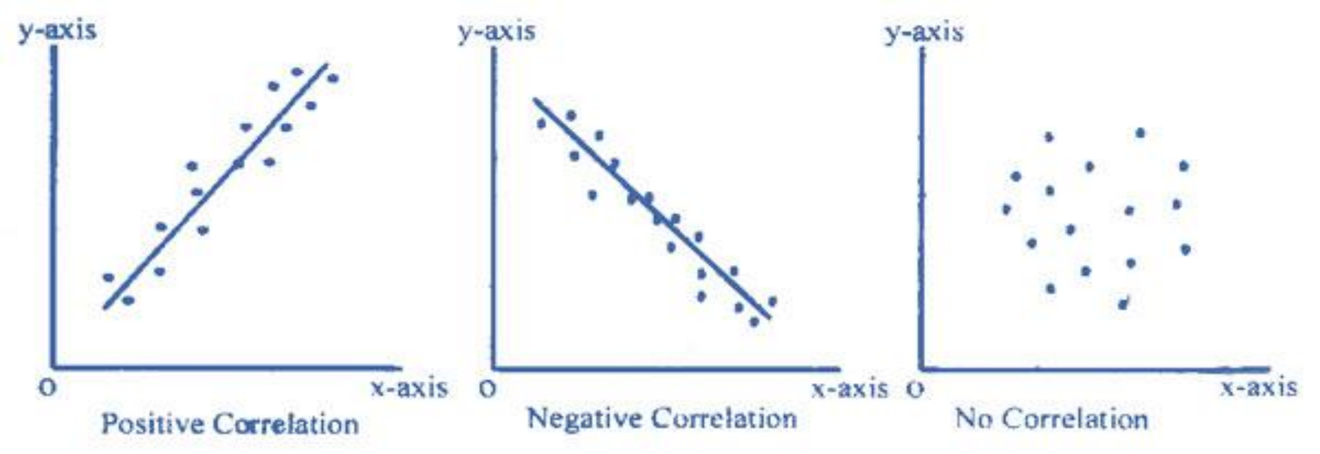
- Intelligence
- Machine Learning
- Human Learning
- AI & Data
- Challenges & Complexities
- AI Cycle
- Applications



Supervised Learning & Linear Regression

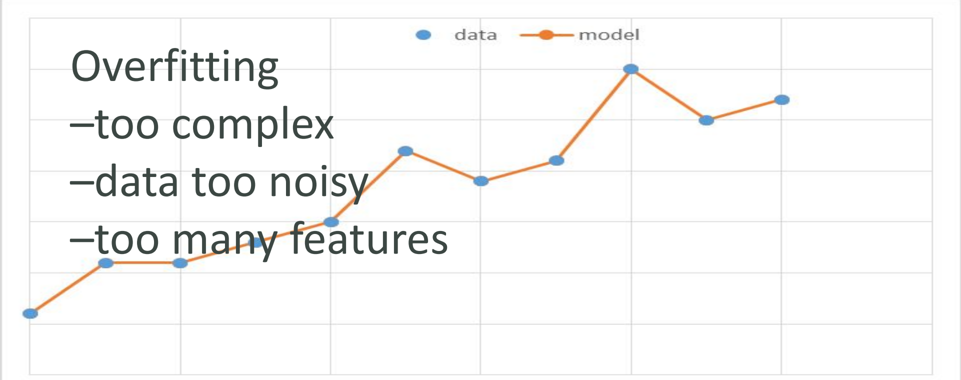
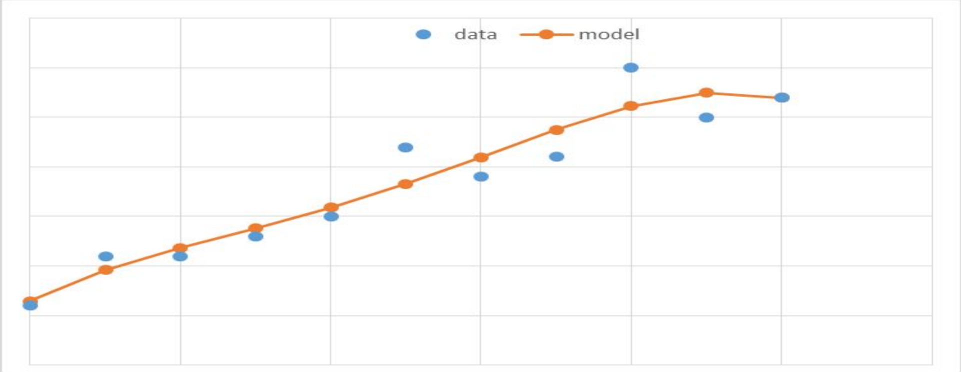
- Types of Machine Learning
- Bias and Variance
- Cost Function & Cross Validation
- Regression vs Classification
- Linear Regression
- Correlation & R2
- Ordinary Least Squares
- Assumptions & Scaling





hi

Bias vs Variance

<table>
<tr><td colspan="2" rowspan="2">DATA FITTING</td><td colspan="2">BIAS (shift)</td></tr>
<tr><td>High -doesn’t respect data, oversimplifies model</td><td>Low</td></tr>
<tr><td rowspan="2">VARIANCE (spread)</td><td>High –too much emphasis on data, poor prediction</td><td></td><td>Overfitting
–too complex
–data too noisy
–too many features
</td></tr>
<tr><td>Low</td><td>Underfitting
–didn’t learn
–not enough data
–under sophisticated model
</td><td></td></tr>
</table>

Gradient Descent

Start with some m_1, b_1, Keep changing m & b to reduce J until we hopefully end up at a minimum

What can go wrong?

IRL: Batch, Stochastic, Mini-batch

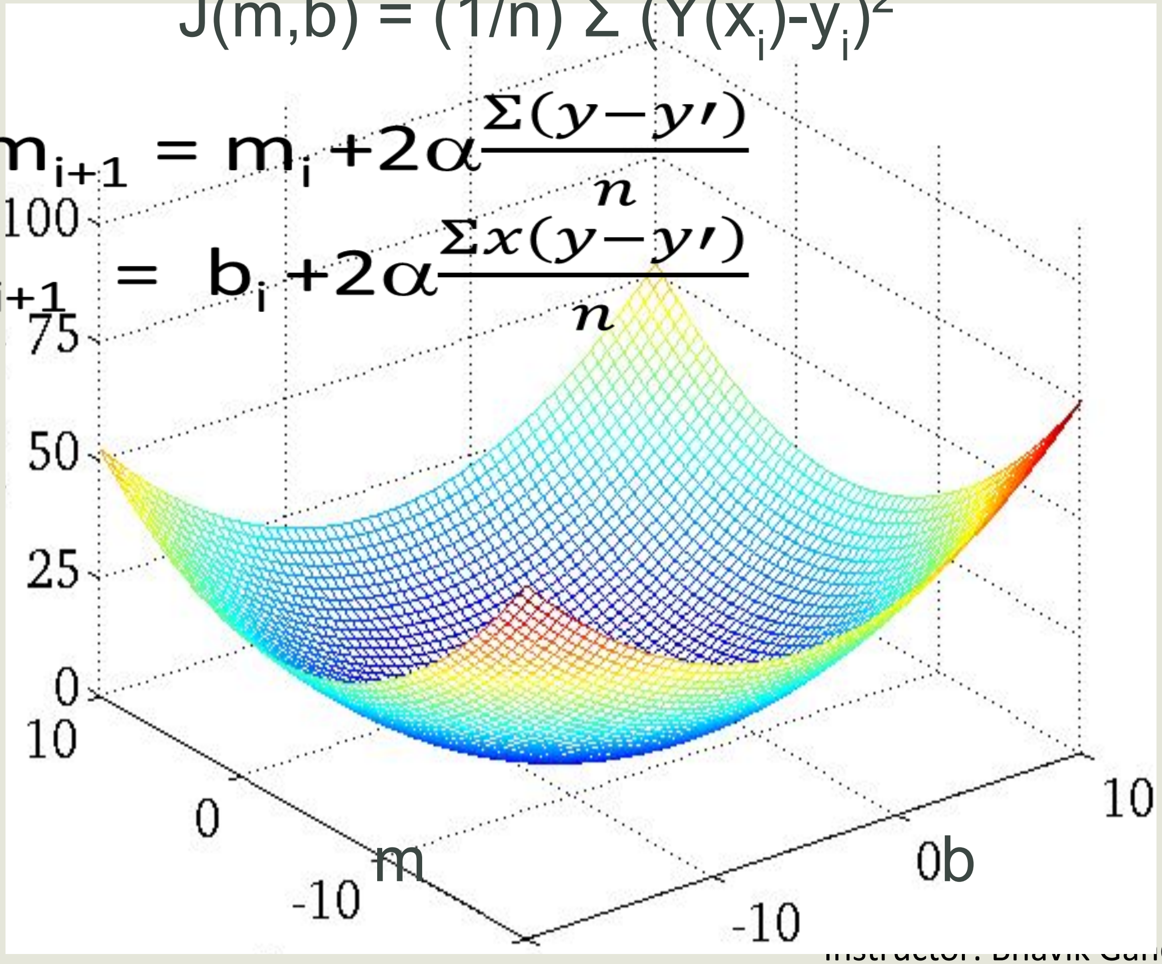
GD vs OLS

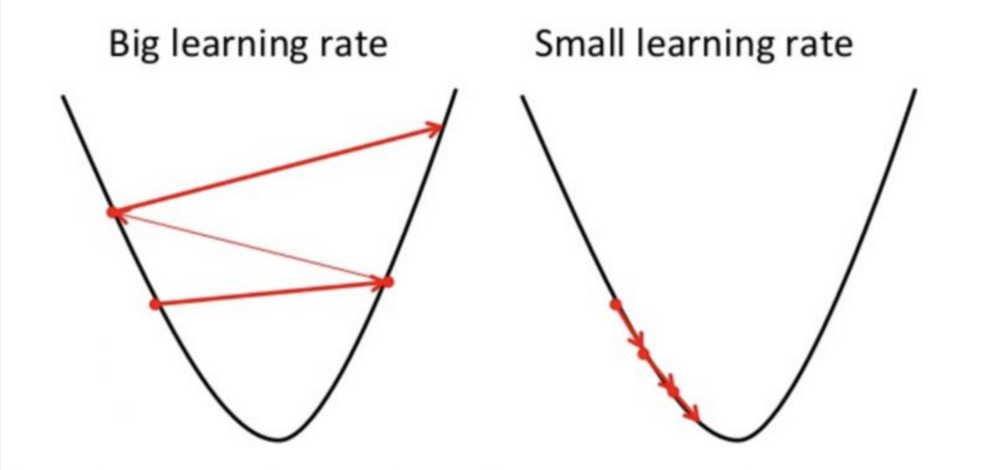
Hyperparameters, Validation

$$J(m,b) = (1/n)\ \Sigma\ (Y(x_i)-y_i)^2$$

$$m_{i+1} = m_i + 2\alpha\frac{\Sigma(y-y')}{n}$$

$$b_{i+1} = b_i + 2\alpha\frac{\Sigma x(y-y')}{n}$$





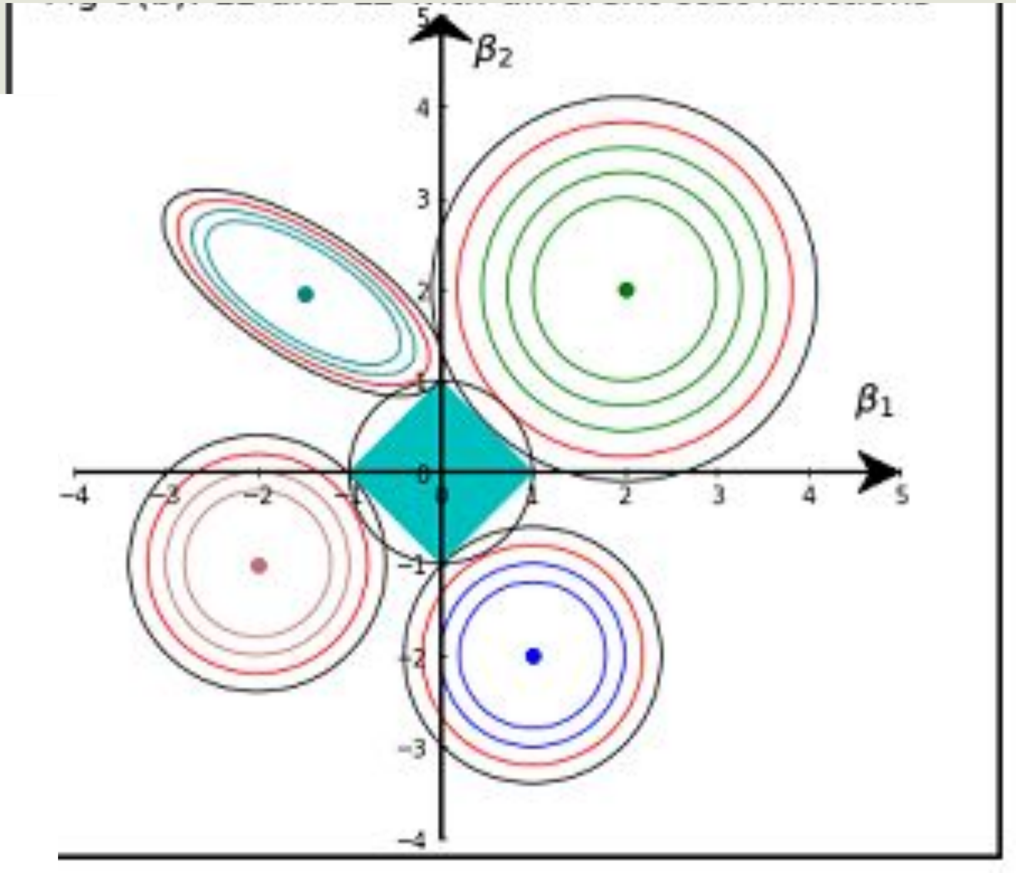
Regularization

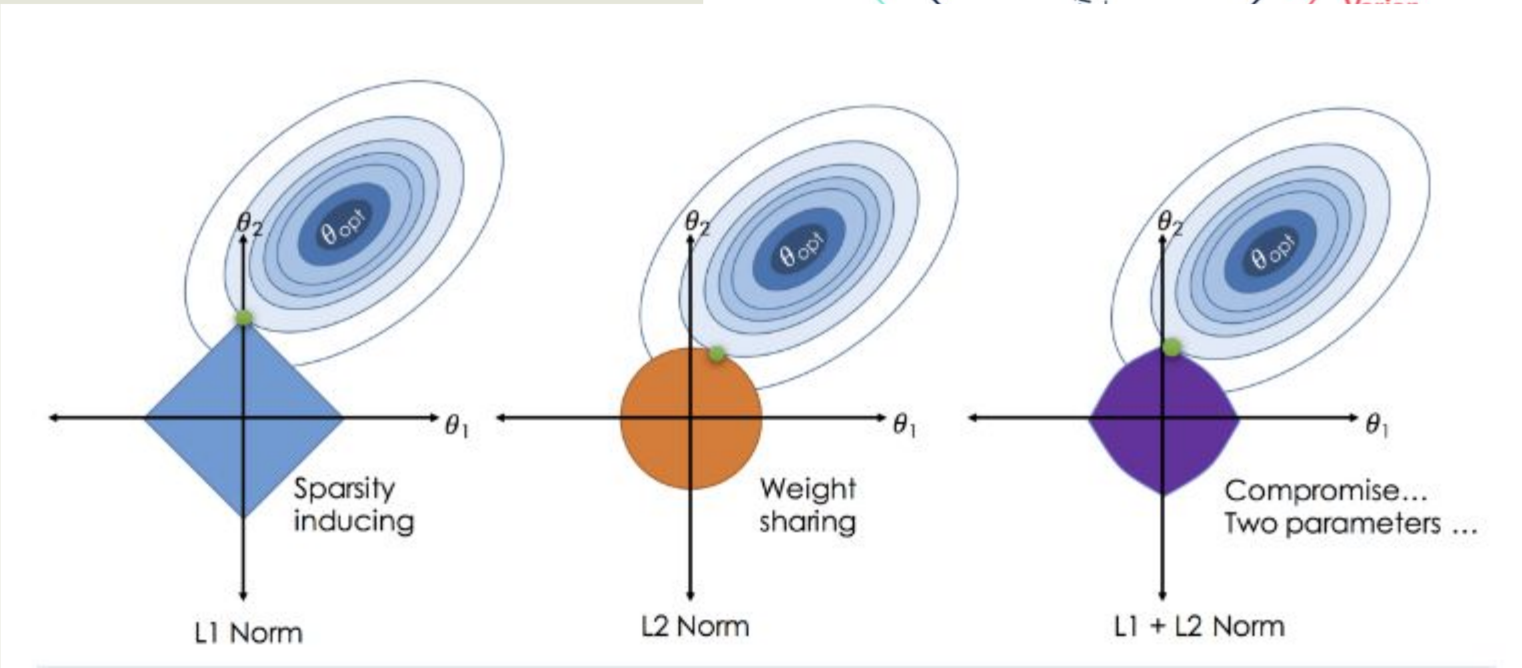
Why do we need Regularization?

What are L1 and L2 norms?

LASSO vs Ridge?

Elastic Net? Downside?





Instructor: Bhavik Gandhi

Logistic Regression

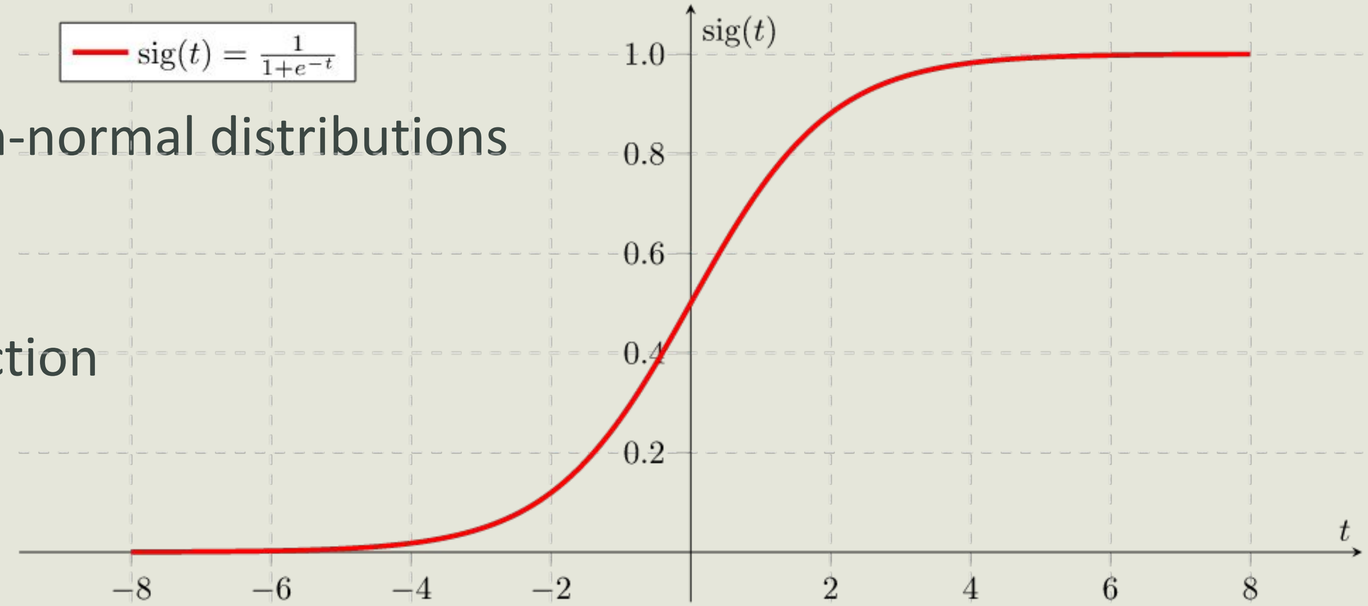
Challenges with using Linear Regression for Classification

- Residuals not normally distributed, not uncorrelated, not homoskedastic, cannot determine Probabilities

Generalized Linear Modelling for non-normal distributions

Use of a Linking function

Using Log-odds and the Sigmoid function



Logistic Regression - Training & Evaluation

Challenges with the Linear Regression Cost Function

Log Likelihood

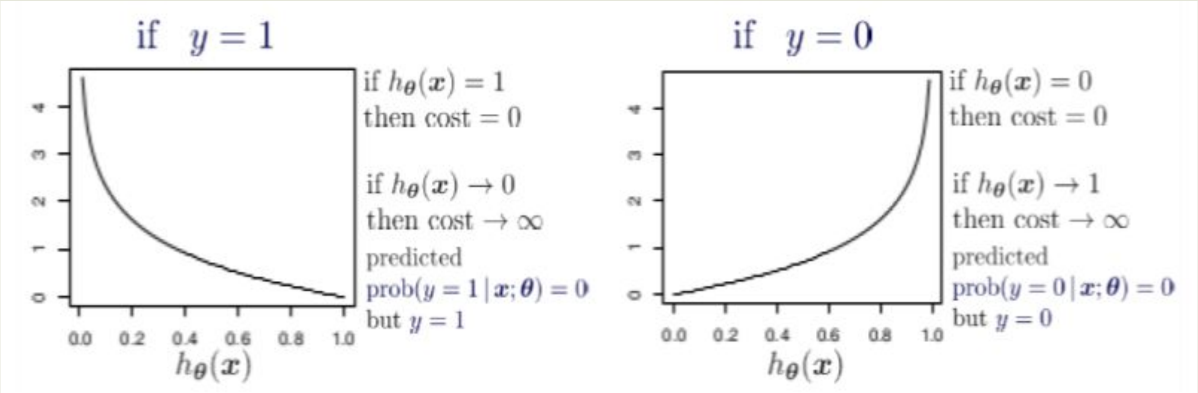
$$-log[L(\theta)] = -\sum_{1}^{n} y * log[\sigma(\theta^T x^i)] + (1-y) * log(1-\sigma(\theta^T x^i)$$

Gradient Descent - Partial Derivatives

$m_{new} = m_{old}$ - alpha * (sigmoid (mx) - y)*x_j

Deviance and Pseudo-R^2

Confusion Matrix



Sensitivity Analysis

What is sensitivity analysis? How does it work?

How does it help? What is it used for?

What are the challenges with SA?

5Qs

5 Questions! Easy-peasy

Shrinkage!

1. Which of these perform Shrinkage?

A. L1 Norm
B. Log Odds
C. L2 Norm
D. Negative Log Loss

Sensitivity Analysis!

2. How many datasets can we have if we have 30 data points and are looking for data sets with 20 data points?

Bias and Variance!

3. Which image (left, center, right) depicts the following?

A. High Variance =
B. High Bias =
C. Good Fit =

Instructor: Bhavik Gandhi

Logistic Regression!

4. What is log-odds? Why do we use log-odds?

Confusion Matrix!

5. For the table alongside

compute the following

A. Precision for Dog Class
B. Recall for Dog Class

Predicted Class	Observed Class		
		Dog	Cat
	Dog	250	100
	Cat	750	900

A. How do you explain the values you got? What do you think is going on?

Answers

1. A & C
2. ${}^{30}C_{20}$ = 30!/20!10! = 29x3x13x5x23x11x21 ~ 30M
3. A = right, B = left, C = center
4. Log-odds is log of odds of success. We use log-odds to achieve an unbounded function just like a linear function
5. A = 0.71, B = 0.25, C = Threshold is very high for Dog Class/Model biased towards Cat Class

Cross Validation

What are we validating?

Machine Learning – Supervised learning

Let’s say we build a model to predict on the data for a given problem, let’s say the model is modelling a function *f* on the data to predict
But how do we know *f* is any good, or that this f is the best?

Typically we define a cost/loss/objective function and measure a ‘price’ paid in doing the fitting/predicting. We usually want to optimize that function.

There are two main ways to deal with this price.

1. Empirically: we assume the data is good, and we want to fit it nicely

But is this the best approach?

Machine Learning – Supervised learning

Goal: minimize loss

1. Empirically: we assume the data is good, and we want to fit it nicely

-but variance is high (model is complex)

What if the data is wrong due to:

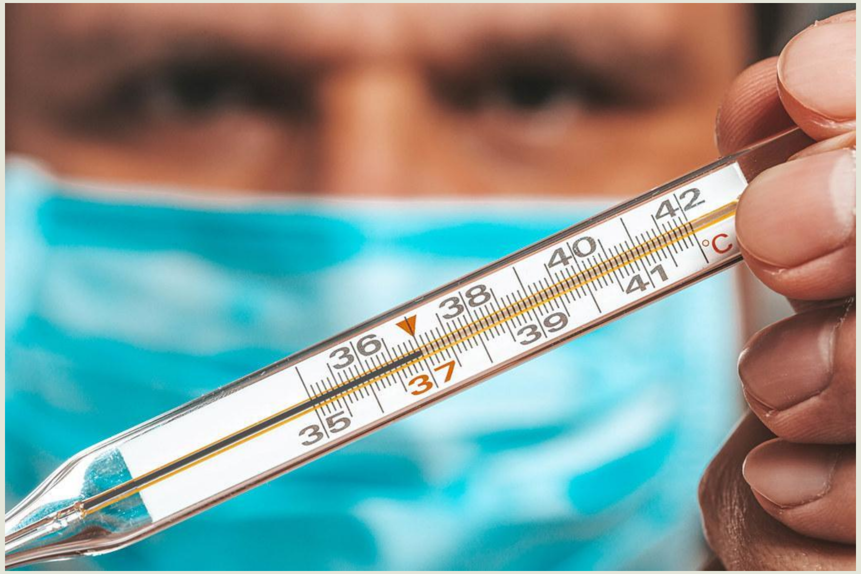
Precision? 37? 37.1? 36.9? 36.899?...

Noise? Well, what is noise?

Erroneous Data?

Missing features?

Is real life continuous/discrete?



What if we are out of our domain and want our model to predict?

Machine Learning – Supervised learning

Goal: minimize loss

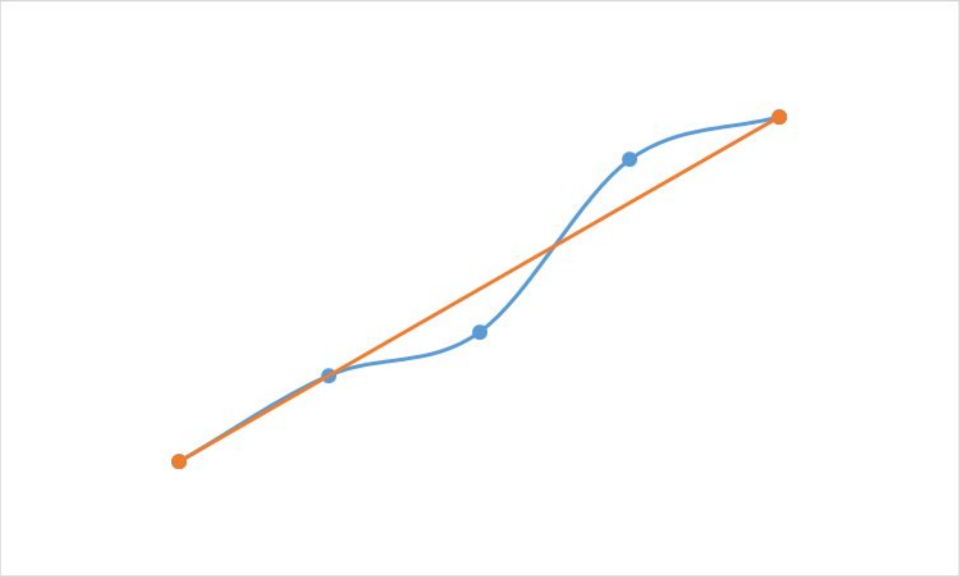
2. Model-wise: we assume the model is good (can predict, smooth, cts.)

But bias is high (systematically shifts outputs/does not fit the data well)

What if the data is great?!

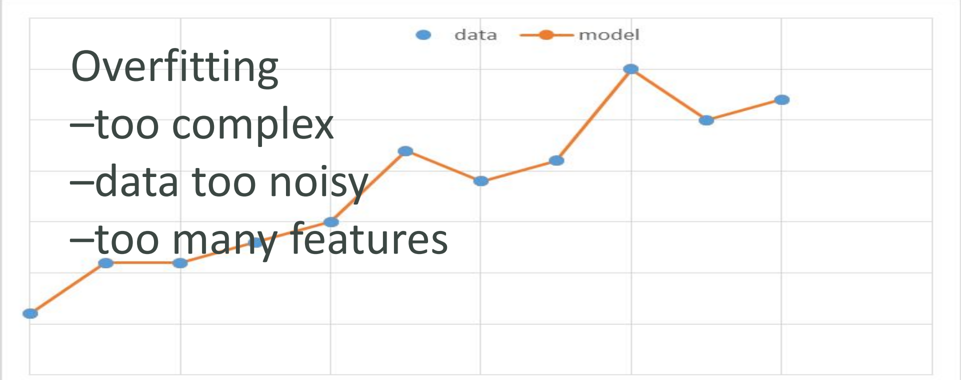
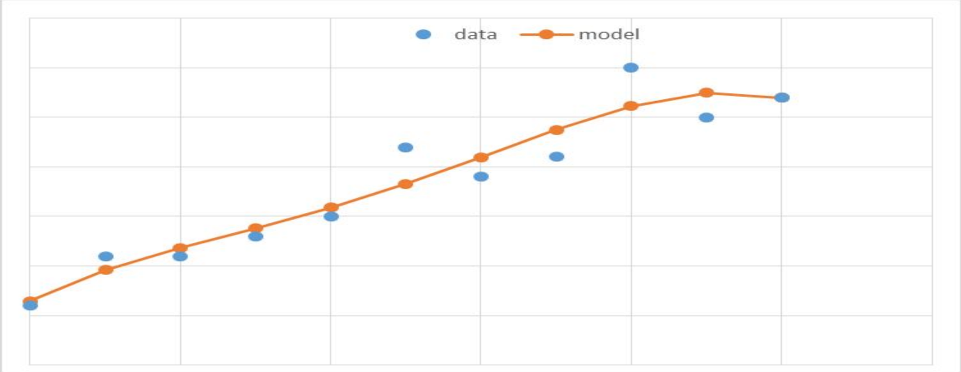
Why did we get it in the first place?

What if the model is too simple?



Machine Learning – Supervised learning

Goal: minimize loss

DATA FITTING		BIAS (shift)	
		High -doesn't respect data, oversimplifies model	Low
VARIANCE (spread)	High –too much emphasis on data, poor prediction		 Overfitting –too complex –data too noisy –too many features
	Low	 Underfitting –didn't learn –not enough data –under sophisticated model	

Machine Learning – Supervised learning

Goal: minimize loss

We want a compromise!

3. Structurally:

We want to design a loss function that attaches importance to low bias and low variance.

In other words we want to control the trade off:

Good Loss ≡ variance + bias

(we want an overall smaller total)

Machine Learning – Supervised learning

Goal: minimize loss by compromising (balance fitting)

3. Structurally:

Loss ~ Exact Interpolation vs. Roughness

So we create a penalty function, *J*,

$$J = \text{RMSE} + \lambda \text{ Regularization(weights)}$$

$\lambda = 0$? We penalize data infidelity ⟶ end up loyal to data (overfit)

$\lambda >> 0$? We penalize roughness ⟶ smooth fit (underfit)

So how do we pick the right λ?

Machine Learning – Supervised learning

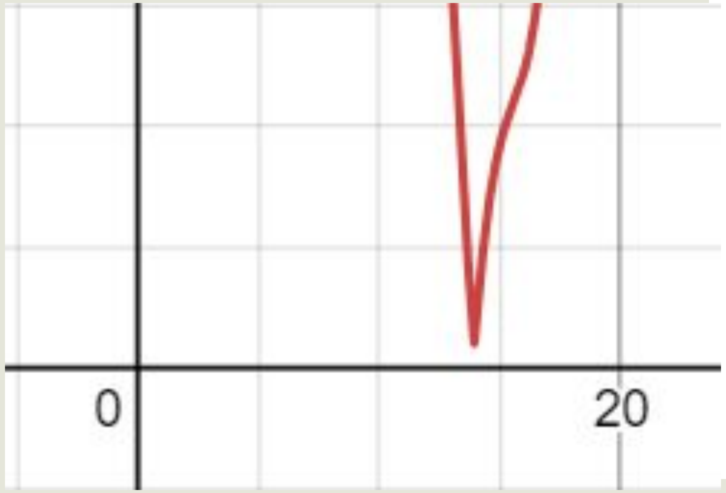
Goal: minimize loss by compromising (balance fitting)

$$J(\lambda) = RMSE + \lambda Regularization(weights)$$

$\lambda = 0 \longrightarrow J = RMSE$

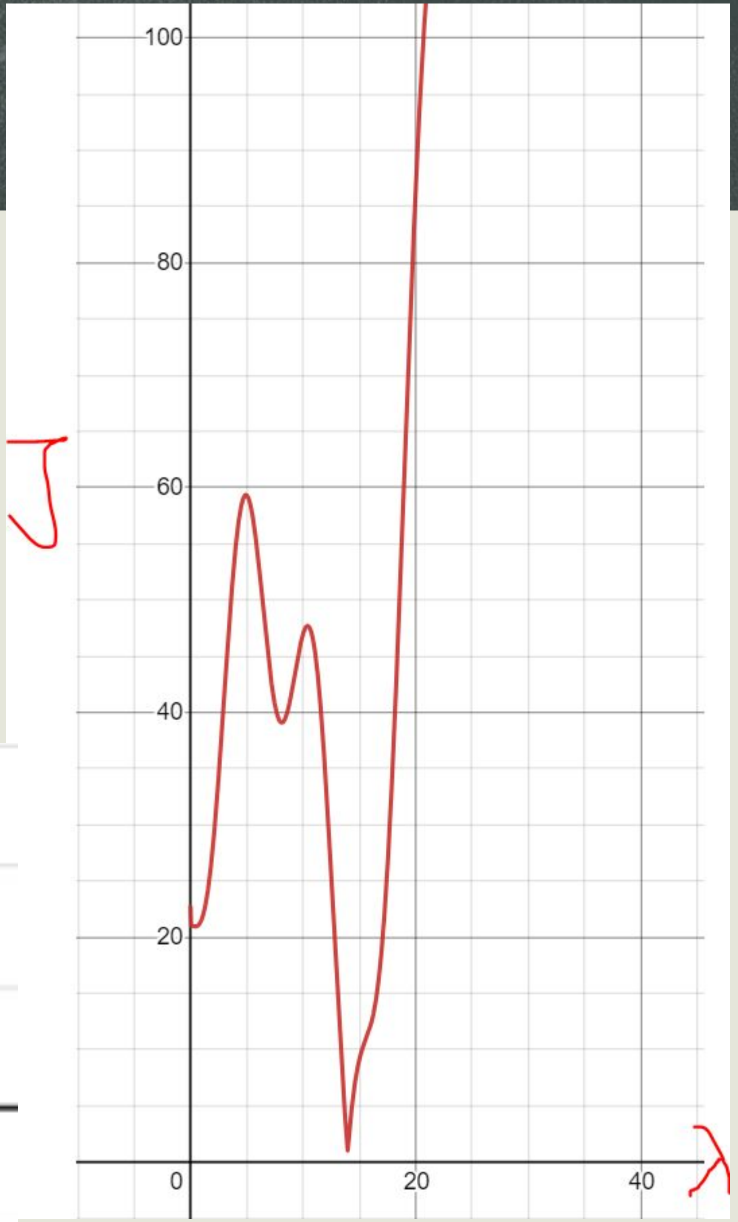
$\lambda >> 0 \longrightarrow J = R_o$

But how do we find the right value of λ



We call this GCV:

Generalized Cross Validation



Hyperparameters

But there is a problem with using our data exclusively.
We have fit our model. We may even have found a good fit.
But after that, then what?!

We have no predictive power.

There are other methods to determine goodness of fit:
One method is to train our model with only some of our data, then see how well it predicts the rest of our data.

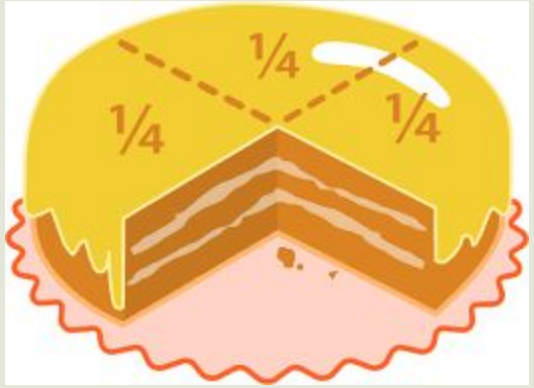
Validation

We usually split our data into three categories that are usually separated:

Training data -the data used to make our initial model

Validation data -the data used to see if our model is good, and to tune it using the method of hyperparametrization

Testing data -the final test of a tuned model is undertaken on this set



Validation

There are many, many different ratios, and schemes that can be used to split the data.

How much data is available?

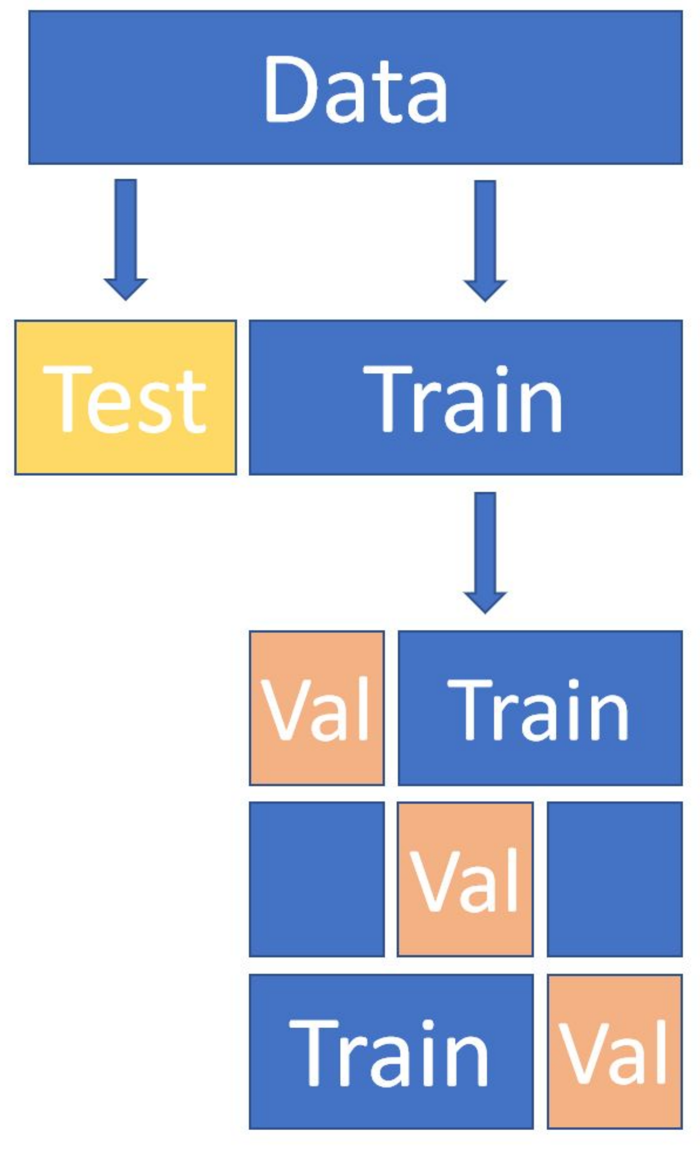
Is it sequential?

Is it all of the same quality?

How many hyperparameters are there?

How sensitive is the model?

How much data does the model need?

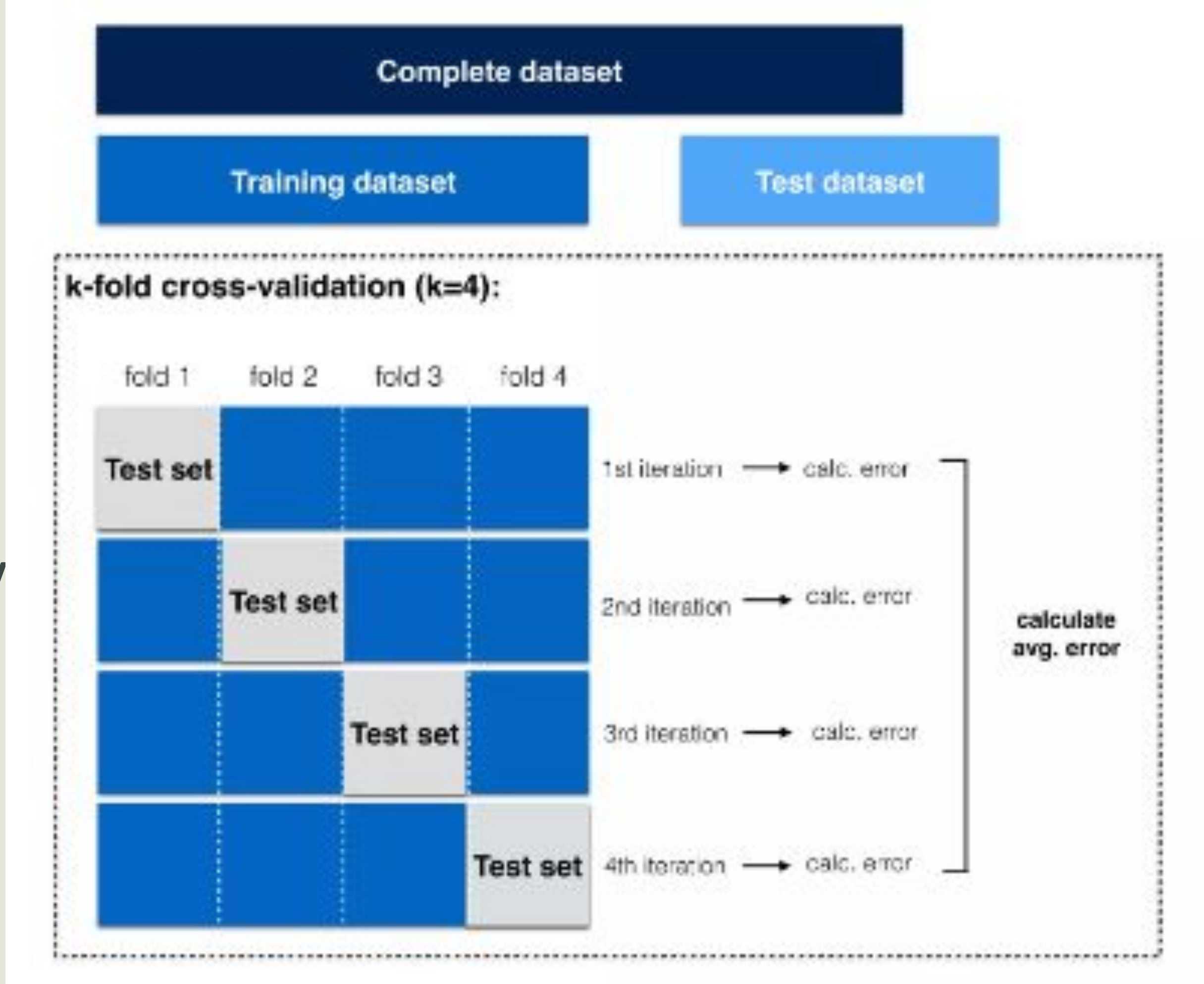


k-fold Cross Validation

Identify average error

Can make use of all the data

You would have more trust and confidence in your model's perf if the accuracy o/ps were 94.0, 92.8, 93.0, 97.0 & 94.5 in a 5-fold cross validation than a 93.0 accuracy in a train-test split. This proves to us that our algo is generalizing, actively learning and is providing consistent reliable o/ps.



Ideal value of k in k-fold

Commonly used values for k are 5 or 10 and there is a good reason for it

Balance using the data with computational effectiveness and test sample size

With k = 10, each sample is 10% or k = 5, each sample is 20% so it’s a 90-10 or 80-20 split, and we’d have 5-10 values of error to get a good average

Also if k is larger the test set will not be as representative and computationally ineffective

Whereas with small k values we’d end up with high bias

Other forms of Cross Validation

Besides k-fold, other types are

1. Hold out

   Same as train-test split

2. Leave one out

   Leave one data point out for testing in each iteration. Not much used currently due to high variance and computational ineffectiveness concerns

3. Leave p-out

   Similar to LOOCV except here p data pts are left out each iteration

4. Rolling CV

   Often used for time-series data, where data from x time is used for training and that from x+n time is used for validation, where n > 0

**Grid search CV, Stratified k-fold CV

Questions

What is k-fold CV? What is a good value for k?

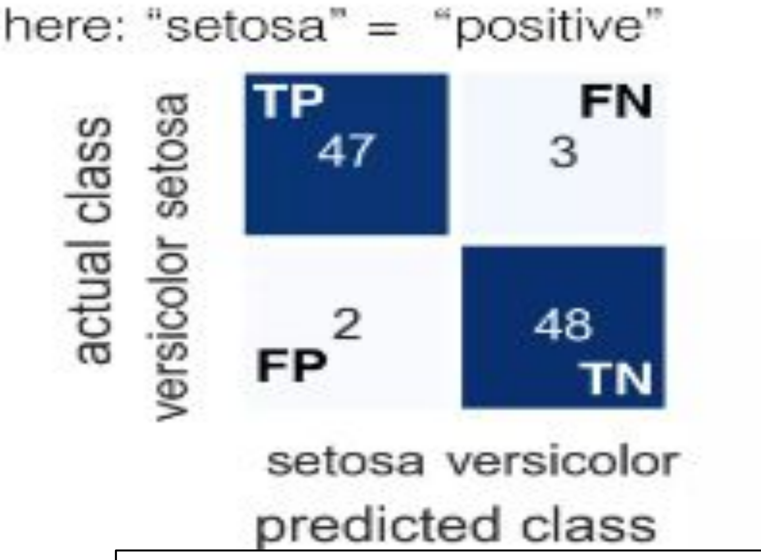
What is the challenge with a 50-fold CV?

What are some other types of CV?

More Evaluation

Haven’t we already done a lot of evaluation in the last couple of weeks? We even had a mid-term

Remember the Confusion Matrix



"micro" and "macro"
averaging for multi-class

$$\textit{Accuracy} = \frac{TP + TN}{FP + FN + TP + TN} = 1 - \textit{Error}$$

$$\textit{False Positive Rate} = \frac{FP}{N}$$

$$\textit{True Positive Rate (Recall)} = \frac{TP}{P}$$

$$\textit{Precision} = \frac{TP}{TP + FP}$$

ROC Curve

- A **receiver operating characteristic curve**, i.e. **ROC curve**, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.
- The diagnostic performance of a test, or the accuracy of a test to discriminate diseased cases from normal cases is evaluated using Receiver Operating Characteristic (ROC) curve analysis

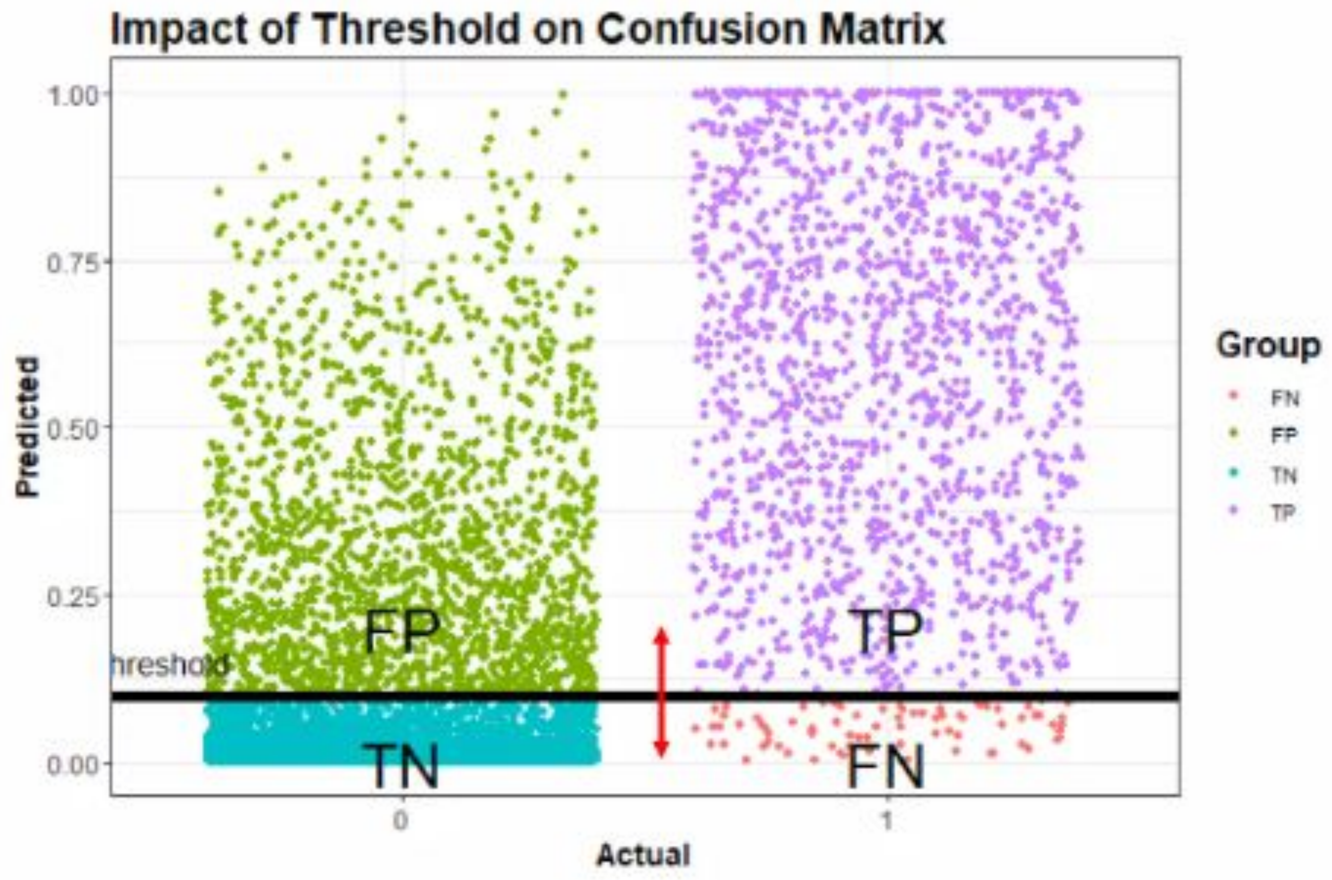
- A Receiver Operating Characteristic (ROC) Curve is a way to compare diagnostic tests. It is a plot of the true positive rate against the false positive rate.

History

Receiver Operating Characteristics (**ROC**) Curve

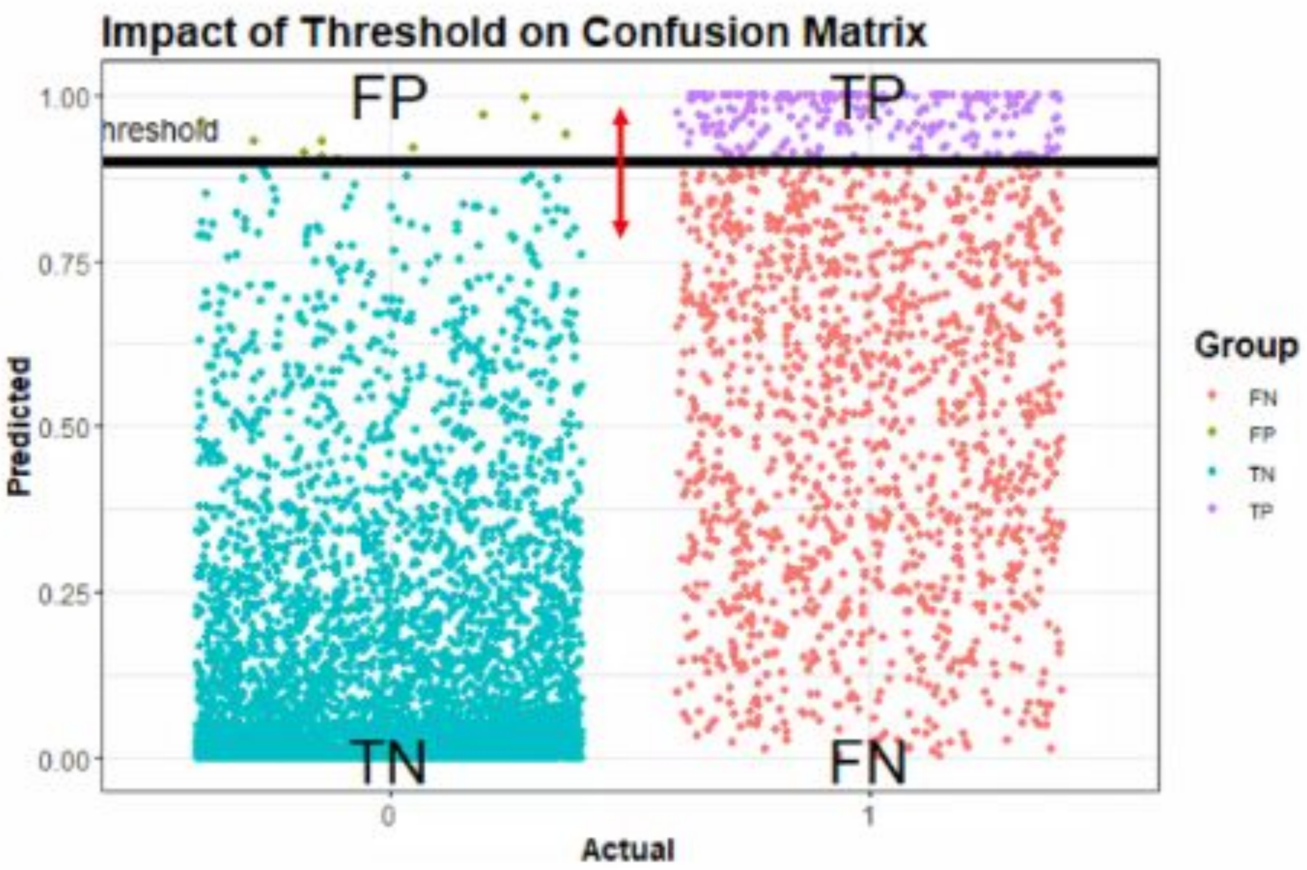
- First developed and used during WWII for detecting enemy objects in battlefields
- Later used in psychology, medicine, forecasting of natural hazards, …
- … and finally **model performance assessment**

Impact of Thresholds



Actuals	PredNeg	PredPos
ActNeg	3117	1842
ActPos	84	1469

Impact of Thresholds



Actuals	PredNeg	PredPos
ActNeg	4948	11
ActPos	1305	248

Instructor: Bhavik Gandhi

From Confusion Matrix to ROC Curve

		Predicted Class	
		Yes	No
Actual Class	Yes	True Pos (Hit)	False Neg (Type I Error)
	No	False Pos (Type II Error)	True Neg (Correct Rejection)

$$TPR = \frac{TP}{TP + FN}$$

→ Y Axis on ROC Curve

		Predicted Class	
		Yes	No
Actual Class	Yes	True Pos (Hit)	False Neg (Type I Error)
	No	False Pos (Type II Error)	True Neg (Correct Rejection)

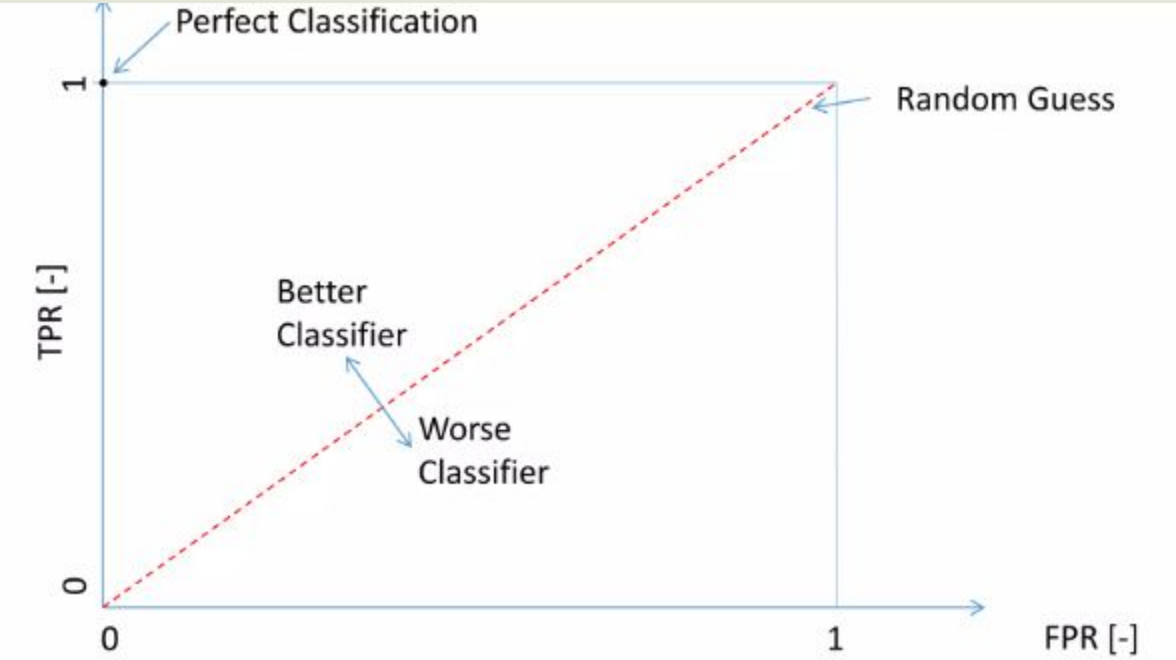
$$FPR = \frac{FP}{FP + TN}$$

→ X Axis on ROC Curve

ROC Curve Example

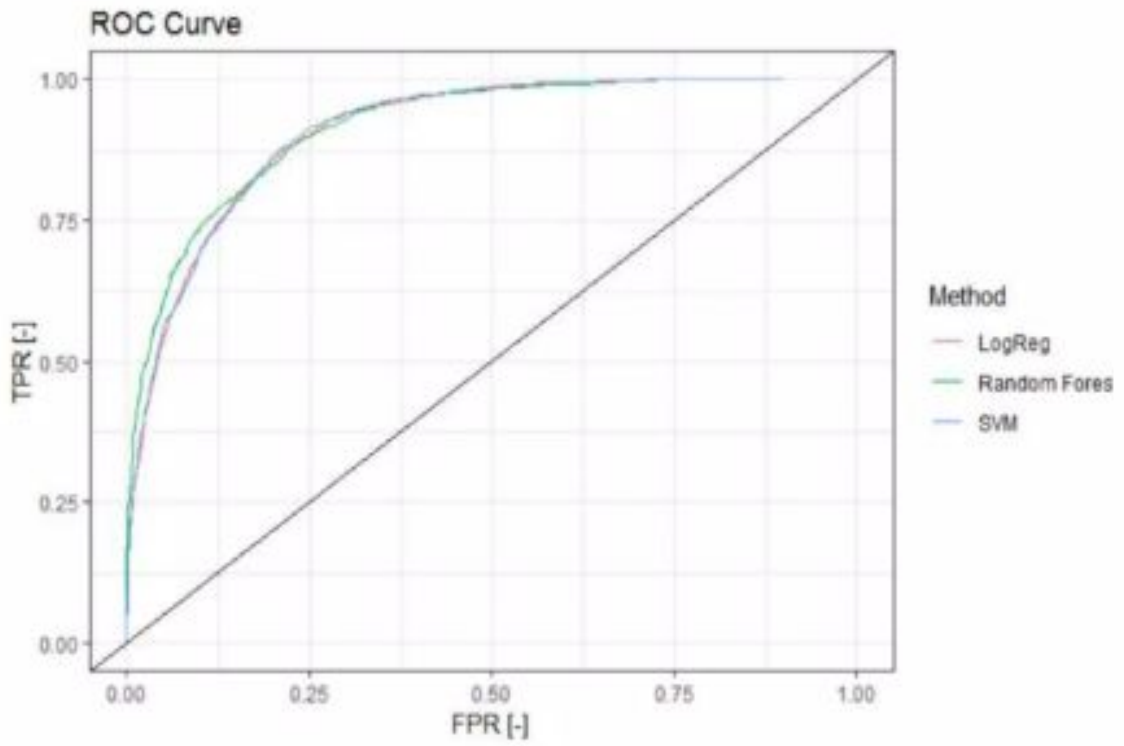
Example

Threshold	TN	FP	FN	TP	FPR	TPR
0.01	1318	3641	3	1550	0.73	1
0.02	1776	3183	10	1543	0.64	0.99
...						
0.98	4958	1	1431	122	0	0.08
0.99	4958	1	1448	105	0	0.07



Instructor: Bhavik Gandhi

Comparing different models

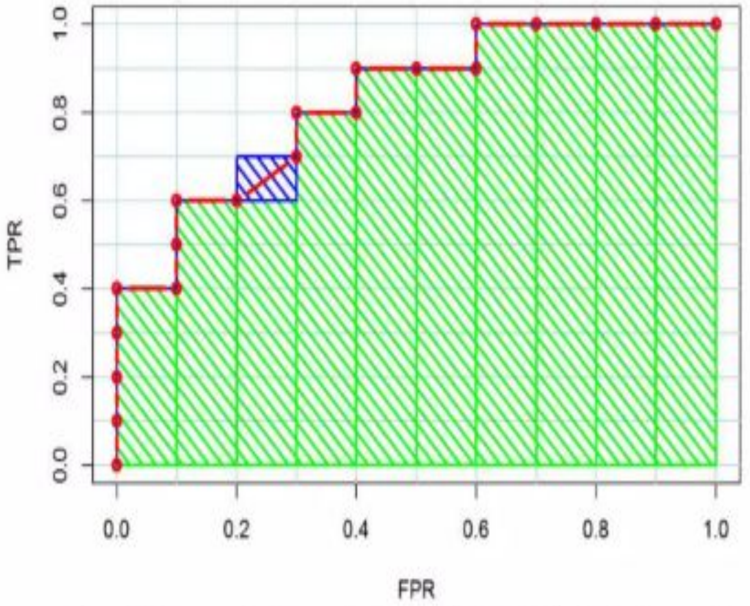


Area under curve

- Maps ROC to one measure
- Purpose: compare different models

Calculates as:

- Sum of green squares
- Half of blue squares



Source: http://blog.revolutionanalytics.com/2016/11/calculating-auc.html

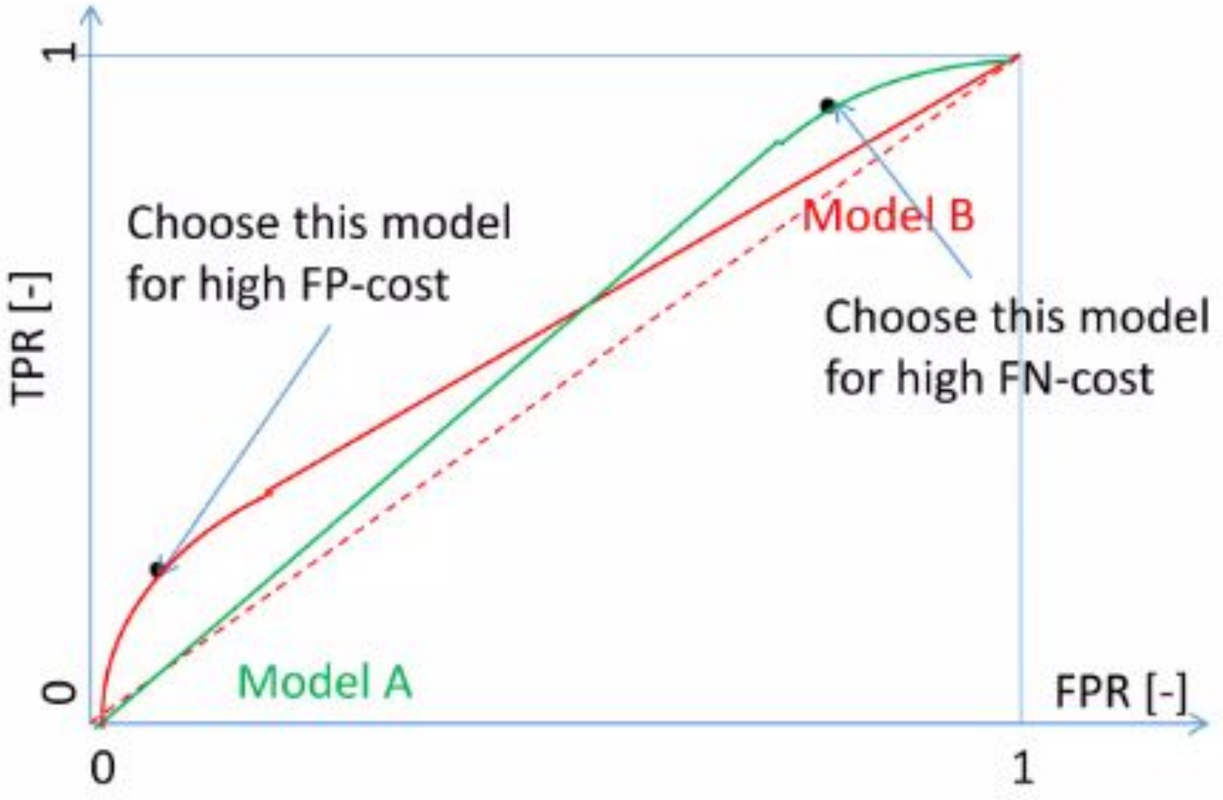
Instructor: Bhavik Gandhi

Choosing between models

- Model A: AUC = 0.6
- Model B: AUC = 0.58
- Which model should you choose? → That depends ☺
- FN and FP might not value identical
- Example: WWII detection of enemy submarine attacks

→ False Negatives (actual attack, predicted no) more critical than False Positive.

- For this you can add different costs, e.g. FN-Cost = 10, FP-Cost = 1



Instructor: Bhavik Gandhi

AUC Challenges

- AUC can be "noisy" (*Small-sample precision of ROC-related estimates*). It targets models that achieve false positive and true positive rates that are significantly above random chance which isn't necessarily good for accuracy.
- A reliable and valid AUC estimate can be interpreted as the probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example.

Alternatives to ROC

Commonly Used F-measures: $F_\beta = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$

- F1_score: $F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2TP}{2TP+FP+FN}$. Also called the balanced score. Can be biased due it being a measure of the mean.
- $F_{0.5}$ Puts more emphasis on the recall.
- F_2 Puts more emphasis on the precision.

Questions

What is ROC? Why do we need it?

What is AUC or AUROC?

How can we use ROC to compare different models?

What are the challenges with AUC?

What are F-measures?

Attendance

I am present :)

Group Assignment!

Ohh yeah, let’s form groups!

Group Assignment - Get Ready

Form your groups and sit with them, be ready to use the terminal at your seat.

Place all digital devices (smart watches, bluetooth earphones, mobiles, laptops, calculators etc.) in your bag. Put your bag under your desk in front of you

Place your id cards along with your group at the edge of the row.

This is a graded assignment and carries 10% of the grade

Read the instructions carefully. Failure to follow them will result in a 0 on the assessment.

And we learnt Cross Validation!!

I will now cross validate everything! :)