https://ttpoll.com/p/618878

COSC 325: Introduction to Machine Learning

Dr. Hector Santos-Villalobos

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Lecture 14: Data Wrangling and Hyperparameter Tuning

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Class Announcements

Homework/Quizzes:

No quiz this week.
Homework #4 due 10/16

Exams:

Next exam 11/21

Course Project:

Midterm Report due 10/27
Teaming issues. Please contact me.

Lectures:

N/A

Review

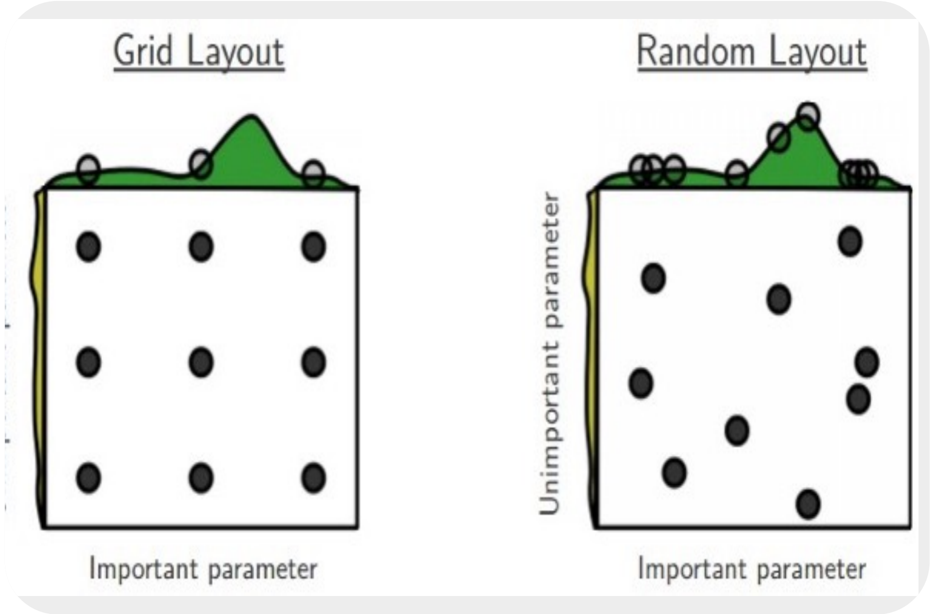
- Ensemble of techniques
- Ensemble of datasets
  - Different datasets -> Bagging
- They Reduce variance
- Perform well as long as only a few of the models make the same mistakes
- Boosting
  - Ensemble of weak learners
  - Easier to design
  - Computational efficient
- Random Forests

https://ttpoll.com/p/618878

Today’s Topics

Data Wrangling

Hyperparameter Tuning



ML Life Cycle

Use Case / Application
- Cancer detection
- Clustering
- Object segmentation
- Control of pressure valve

Machine Learning Category
- Supervised
- Self-supervised
- Semi-supervised
- Reinforcement

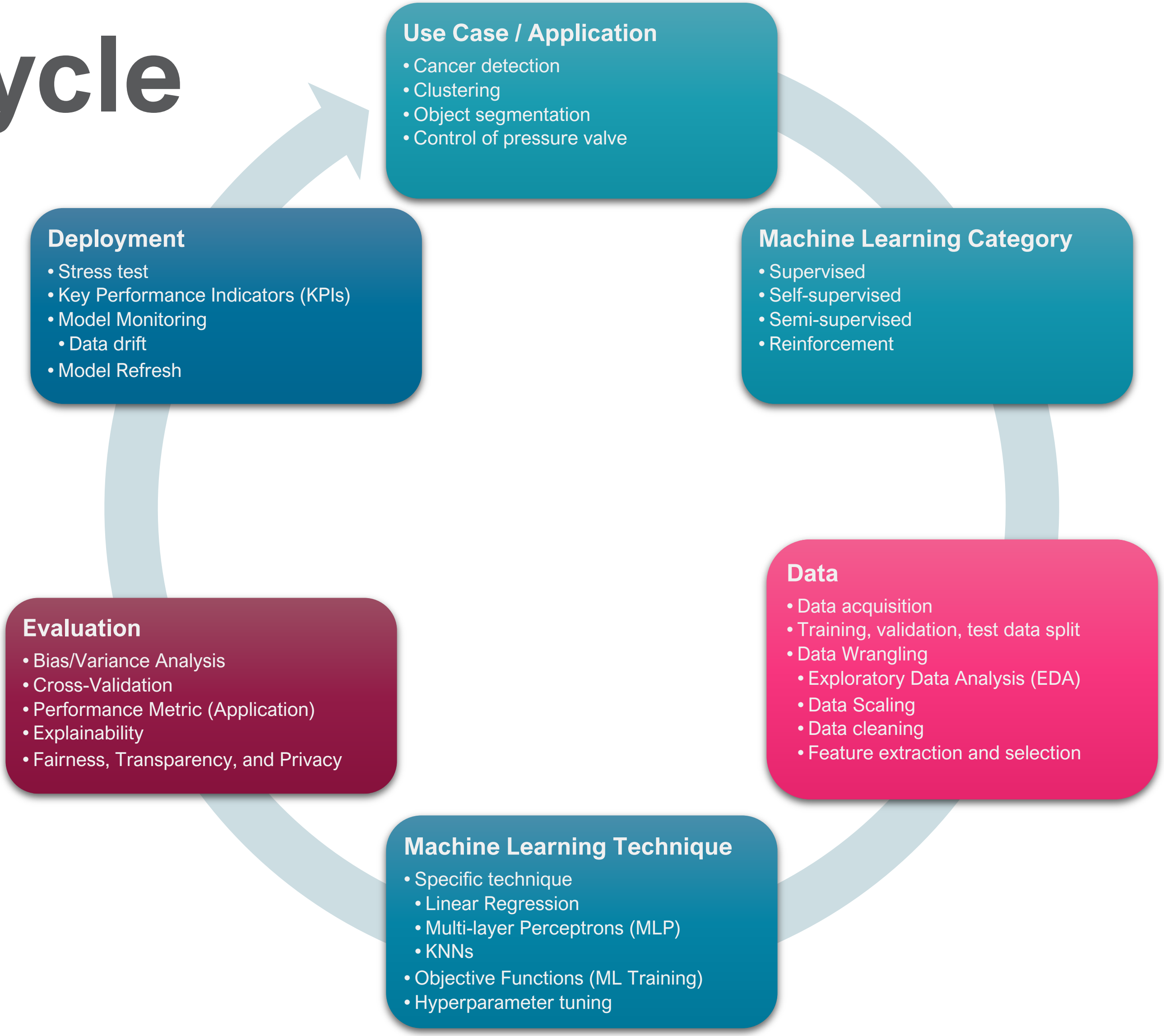
Data
- Data acquisition
- Training, validation, test data split
- Data Wrangling
  - Exploratory Data Analysis (EDA)
  - Data Scaling
  - Data cleaning
  - Feature extraction and selection

Machine Learning Technique
- Specific technique
  - Linear Regression
  - Multi-layer Perceptrons (MLP)
  - KNNs
- Objective Functions (ML Training)
- Hyperparameter tuning

Evaluation
- Bias/Variance Analysis
- Cross-Validation
- Performance Metric (Application)
- Explainability
- Fairness, Transparency, and Privacy

Deployment
- Stress test
- Key Performance Indicators (KPIs)
- Model Monitoring
  - Data drift
- Model Refresh

ML Life Cycle



For in-depth discussion

Dr. Michaela Taufer: COSC 426 - Intro to Data Mining/Analytics

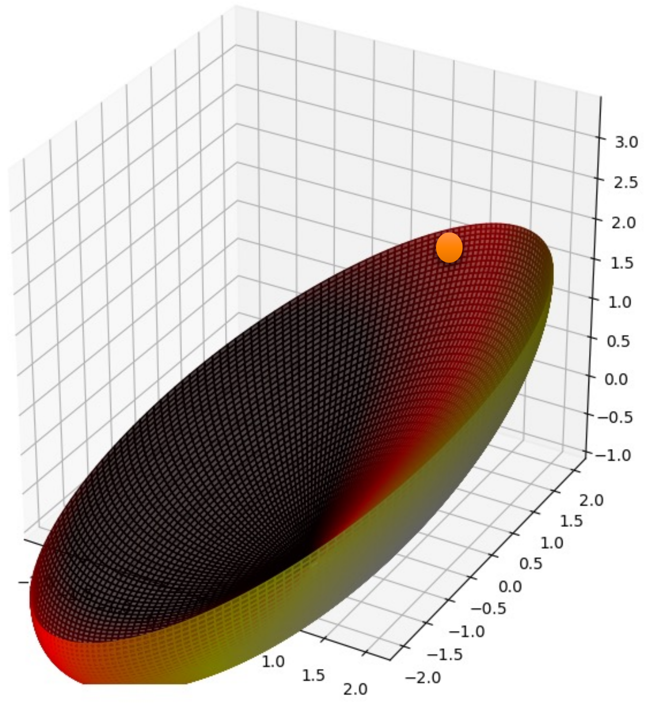
**New name: Data Engineering*

Data Wrangling Topics

- Basic Data Handling
- Preparing Training Data
  - Transformers (Data manipulation/Not DL Technique)
  - Pipelines

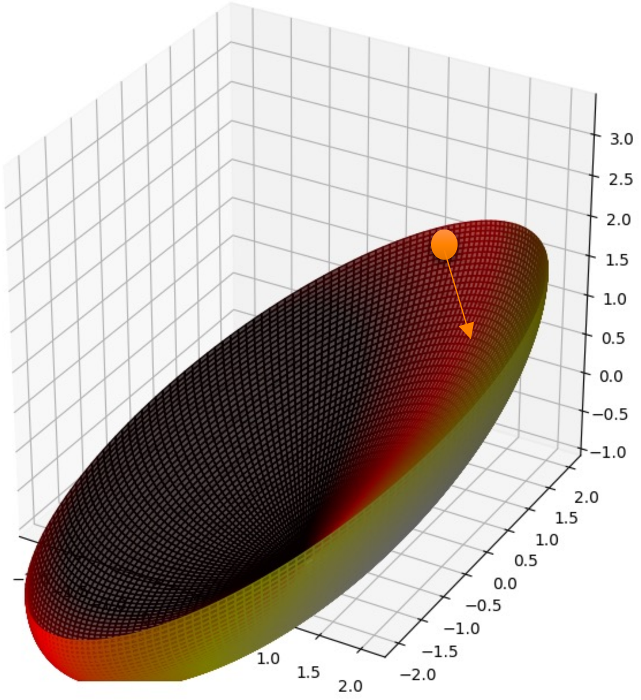
Models like data centered around zero

Models like data with symmetric, low variance.



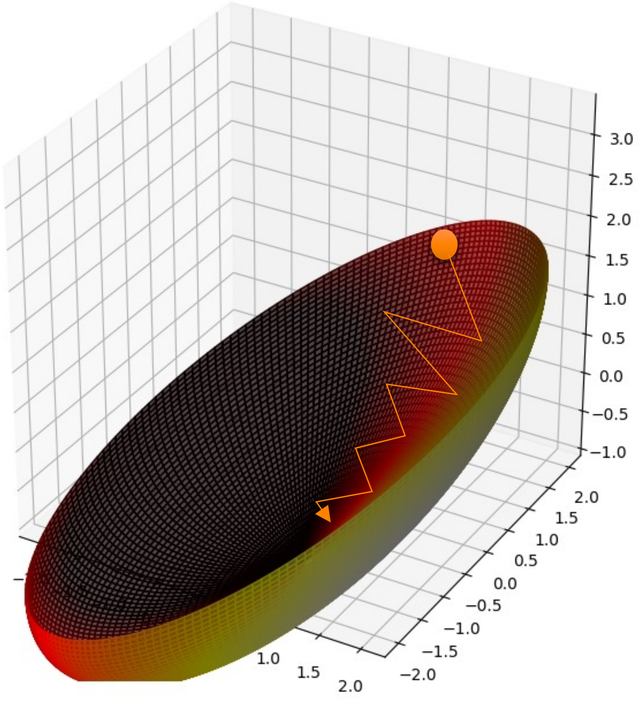
Models like data centered around zero

Models like data with symmetric, low variance.



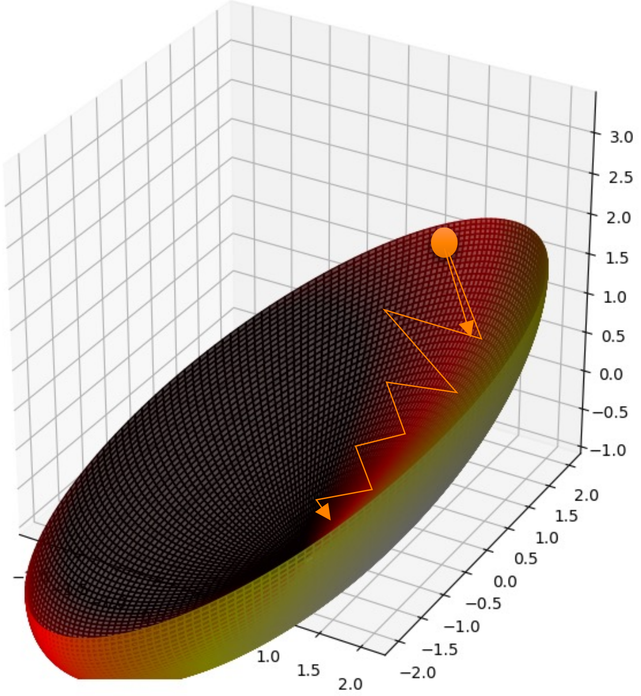
Models like data centered around zero

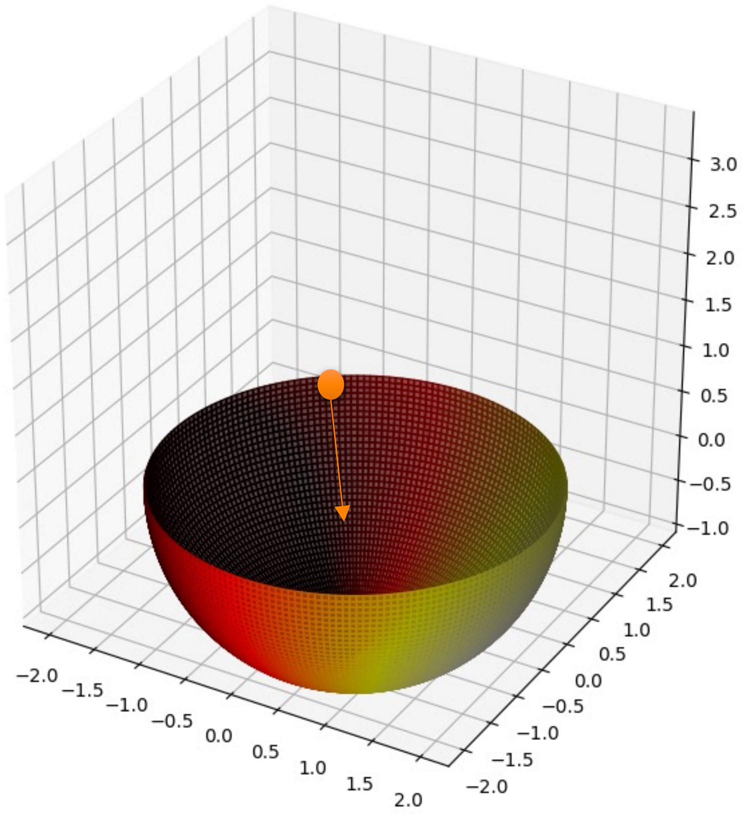
Models like data with symmetric, low variance.



Models like data centered around zero

Models like data with symmetric, low variance.





Normalization

- Give equal weight to all features
- E.g., for a hypothesis $\hat{y} = w_0 + w_1 x_1 + w_2 x_2$,
  - We initialize our weights w_i near zero
  - Then, if $x_2 \gg x_1$ it can take a while for the algorithm to find and appropriate weight w_1 to match the contributions of x_2.
- For gradient-based techniques, normalized inputs prevent too large or too small gradients.

Normalization

- Min-Max: [0,1] range
  - ML technique is sensitive to feature scale (e.g., KNN, SVM, NNs).
  - Data is not normally distributed (e.g., uniform or skewed data).
  - Input features need to be bounded within a specific range
    - E.g., image processing, real-time systems

Normalization

- Min-Max: [0,1] range
  - ML technique is sensitive to feature scale (e.g., KNN, SVM, NNs).
  - Data is not normally distributed (e.g., uniform or skewed data).
  - Input features need to be bounded within a specific range
    - E.g., image processing, real-time systems
- Standardization: Mean μ is zero and standard deviation σ is one.
  - Data follows a Normal distribution
    - Allows comparing the features spread
  - Data variance is more important than the scale.
    - E.g., age in years vs. income in dollars
  - Algorithm assumes data centered around zero (e.g., L2 and L1 Regularization, Neural Networks Tanh activations, etc.)

Min-Max Example

- $x^{(i)} = \frac{(x^{(i)} - x_{min})}{(x_{max} - x_{min})}$

- Training samples:
  - $x^{(1)} = 10\ cm \rightarrow class2$
  - $x^{(2)} = 20\ cm \rightarrow class2$
  - $x^{(3)} = 30\ cm \rightarrow class1$

Min-Max Example

- $x^{(i)} = \frac{(x^{(i)} - x_{min})}{(x_{max} - x_{min})}$

$x_{min} = 10$
$x_{max} = 30$

- Training samples:
  - $x^{(1)} = 10\ cm \rightarrow class2$
  - $x^{(2)} = 20\ cm \rightarrow class2$
  - $x^{(3)} = 30\ cm \rightarrow class1$

Min-Max Example

- $x^{(i)} = \frac{(x^{(i)} - x_{min})}{(x_{max} - x_{min})}$

- Training samples:
  - $x^{(1)} = 10\ cm \rightarrow class2$
  - $x^{(2)} = 20\ cm \rightarrow class2$
  - $x^{(3)} = 30\ cm \rightarrow class1$

Normalized →

$$x_{min} = 10$$
$$x_{max} = 30$$

$$x^{(1)} = \frac{10 - 10}{30 - 10} = 0$$

$$x^{(2)} = \frac{20 - 10}{30 - 10} = \frac{10}{20} = 0.5$$

$$x^{(3)} = \frac{30 - 10}{30 - 10} = \frac{20}{20} = 1.0$$

Standardization Example

- $x^{(i)} = \frac{(x^{(i)} - \mu_x)}{\sigma_x}$

$$\mu_x = \frac{1}{n}\sum_i x^{(i)} = \frac{1}{3}(10 + 20 + 30) = 20$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_i (x^{(i)} - \mu_x)^2} = 10$$

- Training samples:
  - $x^{(1)} = 10\ cm \ \rightarrow class2$
  - $x^{(2)} = 20\ cm \ \rightarrow class2$
  - $x^{(3)} = 30\ cm \ \rightarrow class1$

Standardization Example

- $x^{(i)} = \frac{(x^{(i)} - \mu_x)}{\sigma_x}$

$$\mu_x = \frac{1}{n}\sum_i x^{(i)} = \frac{1}{3}(10 + 20 + 30) = 20$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_i (x^{(i)} - \mu_x)^2} = 10$$

Sample Deviation

Population deviation σ_x divides by n instead of $n-1$

- Training samples:
  - $x^{(1)} = 10\ cm \rightarrow class2$
  - $x^{(2)} = 20\ cm \rightarrow class2$
  - $x^{(3)} = 30\ cm \rightarrow class1$

In Numpy: `np.std(x, ddof=1)`

Standardization Example

- $x^{(i)} = \frac{(x^{(i)} - \mu_x)}{\sigma_x}$

- Training samples:
  - $x^{(1)} = 10\ cm \rightarrow class2$
  - $x^{(2)} = 20\ cm \rightarrow class2$
  - $x^{(3)} = 30\ cm \rightarrow class1$

$$\mu_x = \frac{1}{n}\sum_i x^{(i)} = \frac{1}{3}(10 + 20 + 30) = 20$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_i (x^{(i)} - \mu_x)^2} = 10$$

Normalized →

$$x^{(1)} = \frac{10 - 20}{10} = -1.0$$

$$x^{(2)} = \frac{20 - 20}{10} = 0$$

$$x^{(3)} = \frac{30 - 20}{10} = 1.0$$

What about validation and test sets?

- From the training set
  - $\mu_x = 20$
  - $s_x = 10$
- Standardization of Validation samples:
  - $x_v^{(1)} = 13\ cm\ \rightarrow class2$
  - $x_v^{(2)} = 15\ cm\ \rightarrow class2$
  - $x_v^{(3)} = 28\ cm\ \rightarrow class1$

What about validation and test sets?

- From the training set
  - $\mu_x = 20$
  - $s_x = 10$
- Standardization of Validation samples:
  - $x_v^{(1)} = 13\ cm \rightarrow class2$
  - $x_v^{(2)} = 15\ cm \rightarrow class2$
  - $x_v^{(3)} = 28\ cm \rightarrow class1$

We use Training set normalization parameters on Validation and Test sets

Normalized

$$x_v^{(1)} = \frac{13 - 20}{10} = -0.7$$

$$x_v^{(2)} = \frac{15 - 20}{10} = -0.5$$

$$x_v^{(3)} = \frac{28 - 20}{10} = 0.8$$

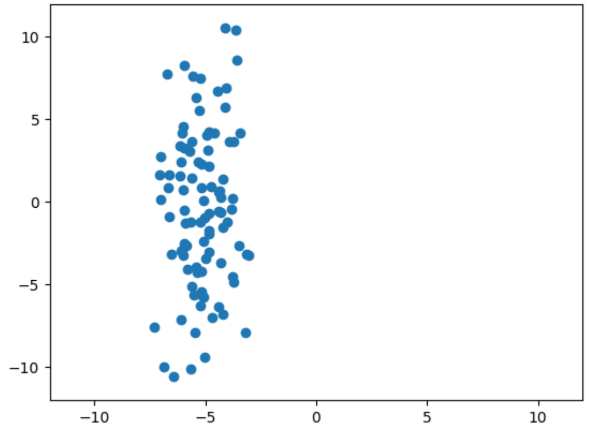
Standardization

$$x = \frac{x_{original} - \mu_x}{\sigma_x}$$

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

$$\sigma_x = \frac{1}{n-1}\sum_{i=1}^{n} (x^{(i)} * x^{(i)})$$

Original Data



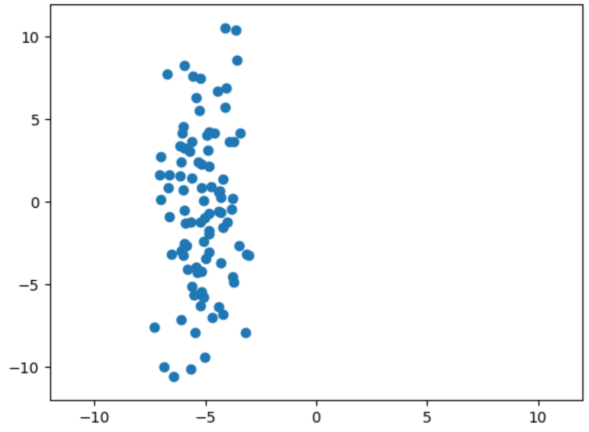
Standardization

$$x = \frac{x_{original} - \mu_x}{\sigma_x}$$

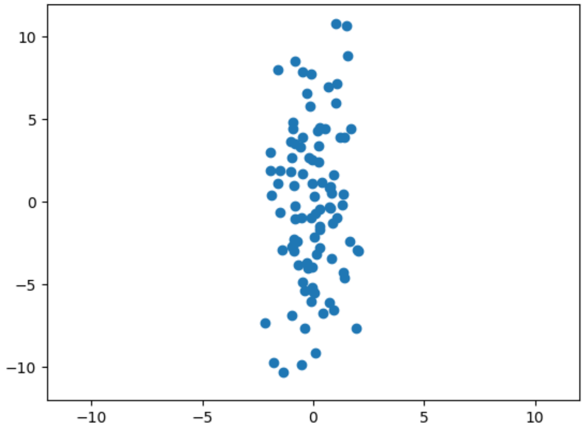
$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

$$\sigma_x = \frac{1}{n-1}\sum_{i=1}^{n}(x^{(i)} * x^{(i)})$$

Original Data



Subtract μ_x



Standardization

$$x = \frac{x_{original} - \mu_x}{\sigma_x}$$

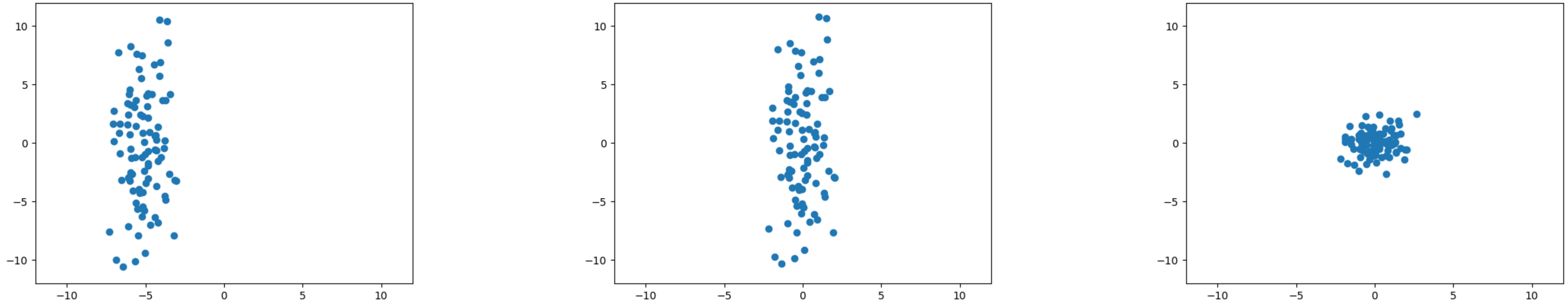
$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

$$\sigma_x = \frac{1}{n-1}\sum_{i=1}^{n}(x^{(i)} * x^{(i)})$$

Original Data

Subtract μ_x

Divide by σ_x



Pop Quiz

True or False. To Min-Max normalize the validation set, we compute the set's minimum x_val_min and maximum x_val_max and apply the formula below to each sample in the set.

x_val_norm=(x_val_sample—x_val_min)/(x_val_max—x_val_min)

A. True

B. False

SkLearn Pipelines

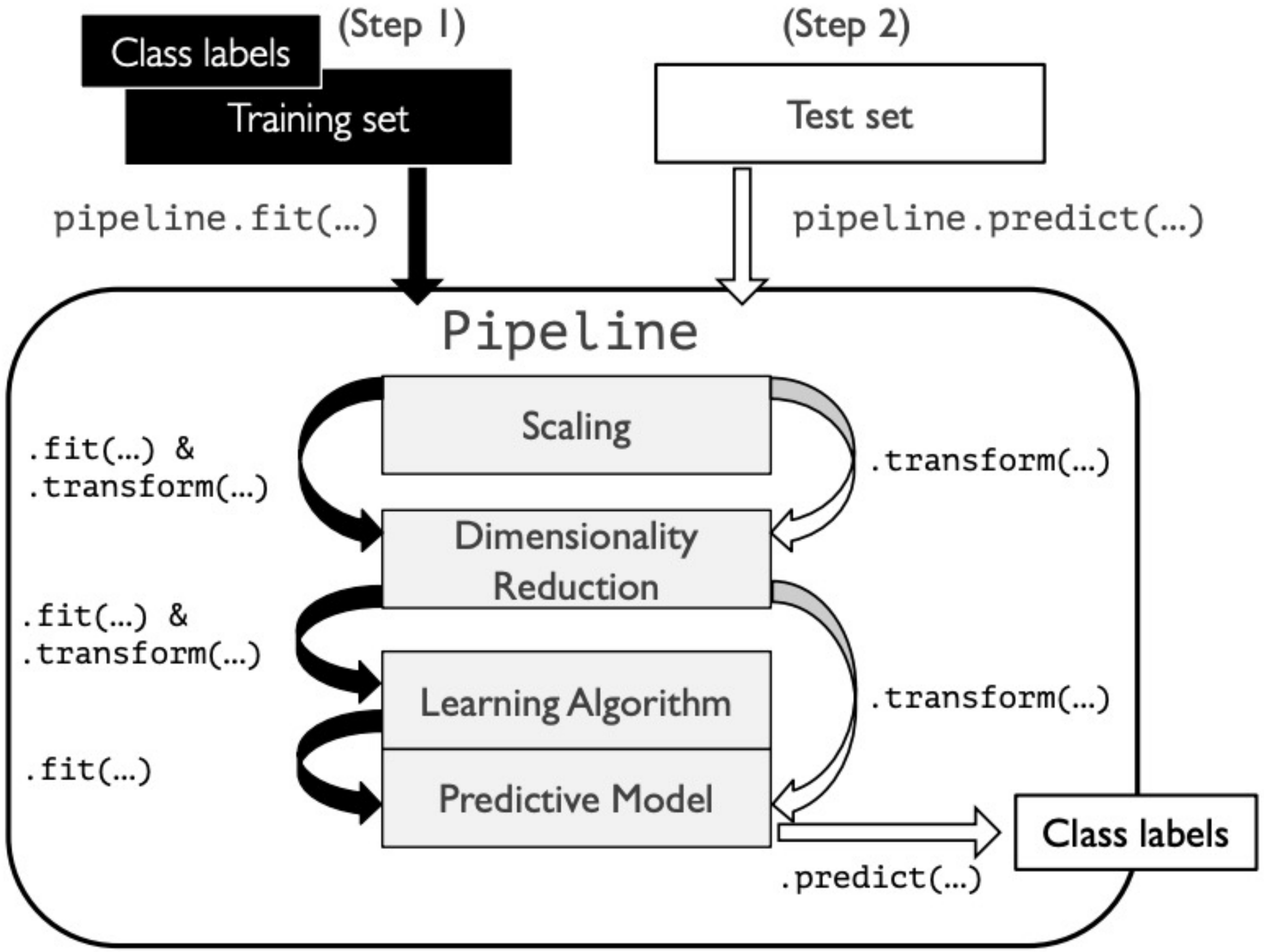


Image source: Dr. Sebastian Raschka,
Machine Learning Course

SkLearn Pipelines

- End-to-end data preprocessing and model training/testing
- Consistent application of data transformations and manipulations
  - Replacing missing values
  - Normalization
  - Encoding
- Avoid skipping steps or using the wrong parameters

Hyperparameter Tuning

THE UNIVERSITY OF TENNESSEE KNOXVILLE

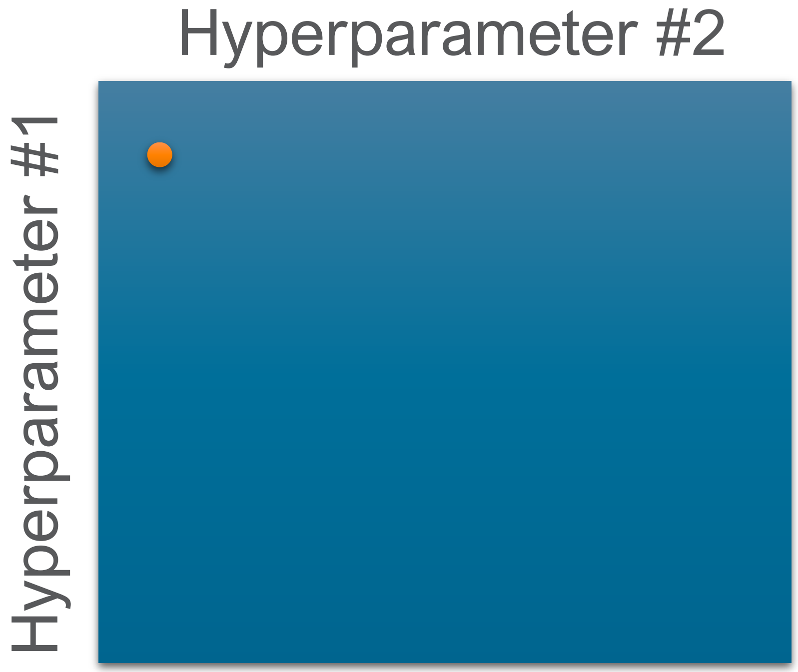
Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

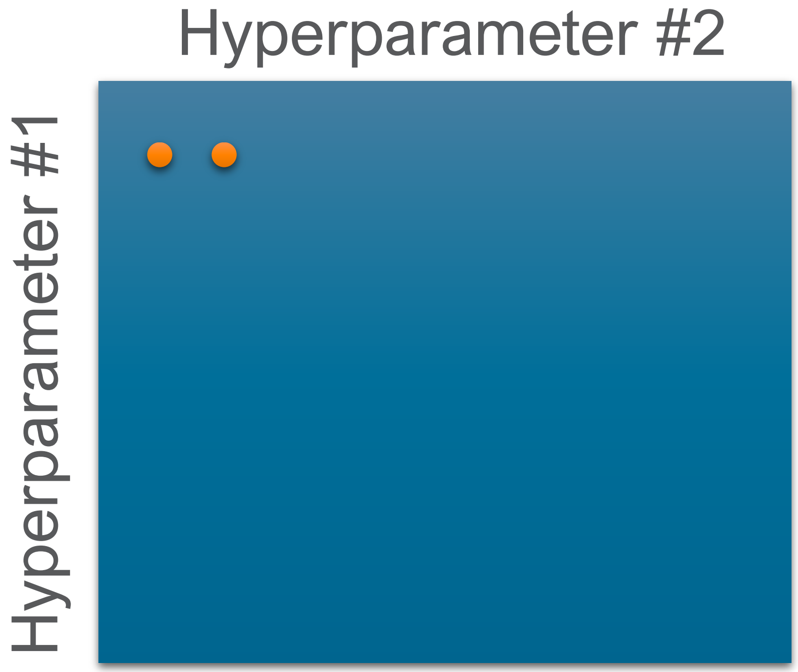
What not to do?



Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

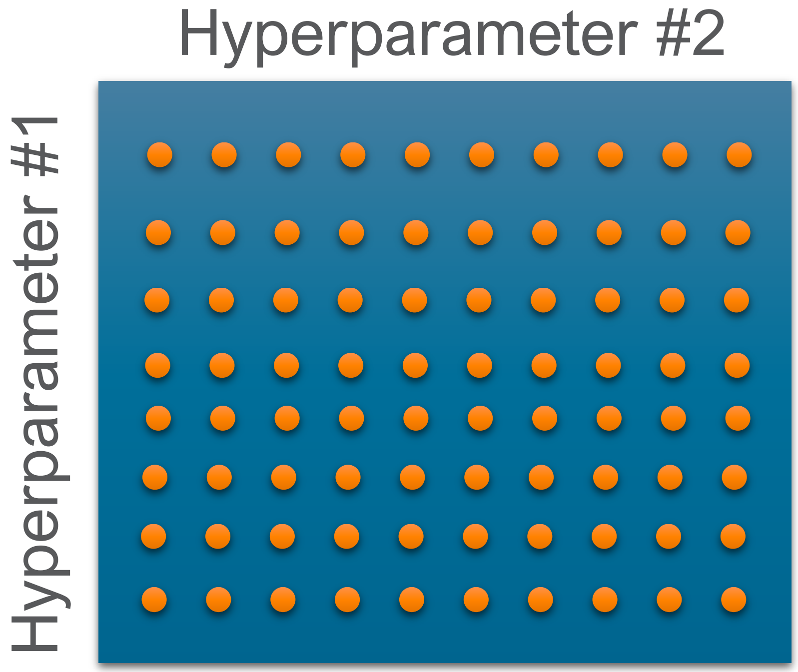
What not to do?



Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

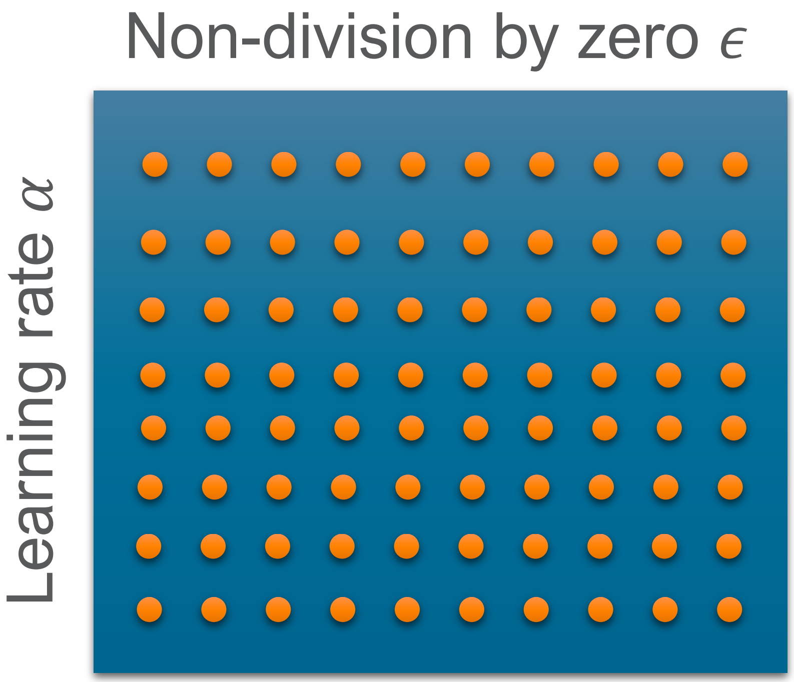
What not to do?



Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

What not to do?

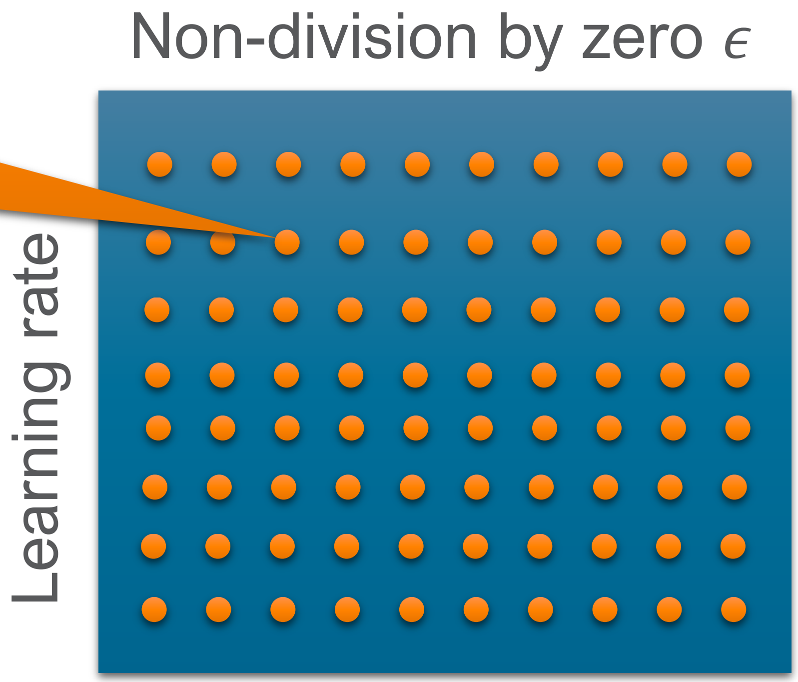


Hyperparameters

- Learning rate α
- Mini-batch
- Decision T
  - Tree dept
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

We will test 10 values of a low priority parameter without changing a high-priority parameter.

<u>What not to do?</u>

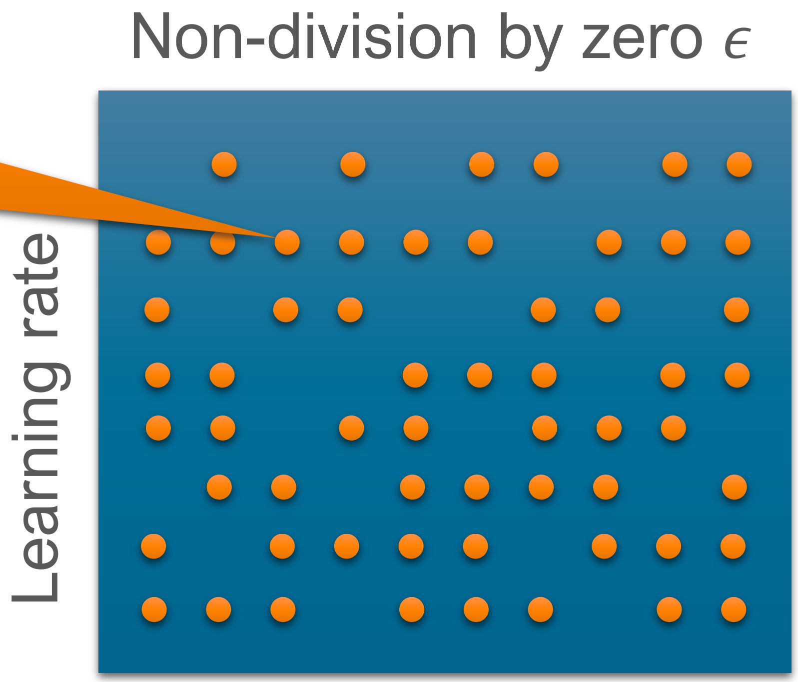


Hyperparameters

- Learning rate α
- Mini-batch
- Decision T
  - Tree dept
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

Randomly sample the hyperparameter space.

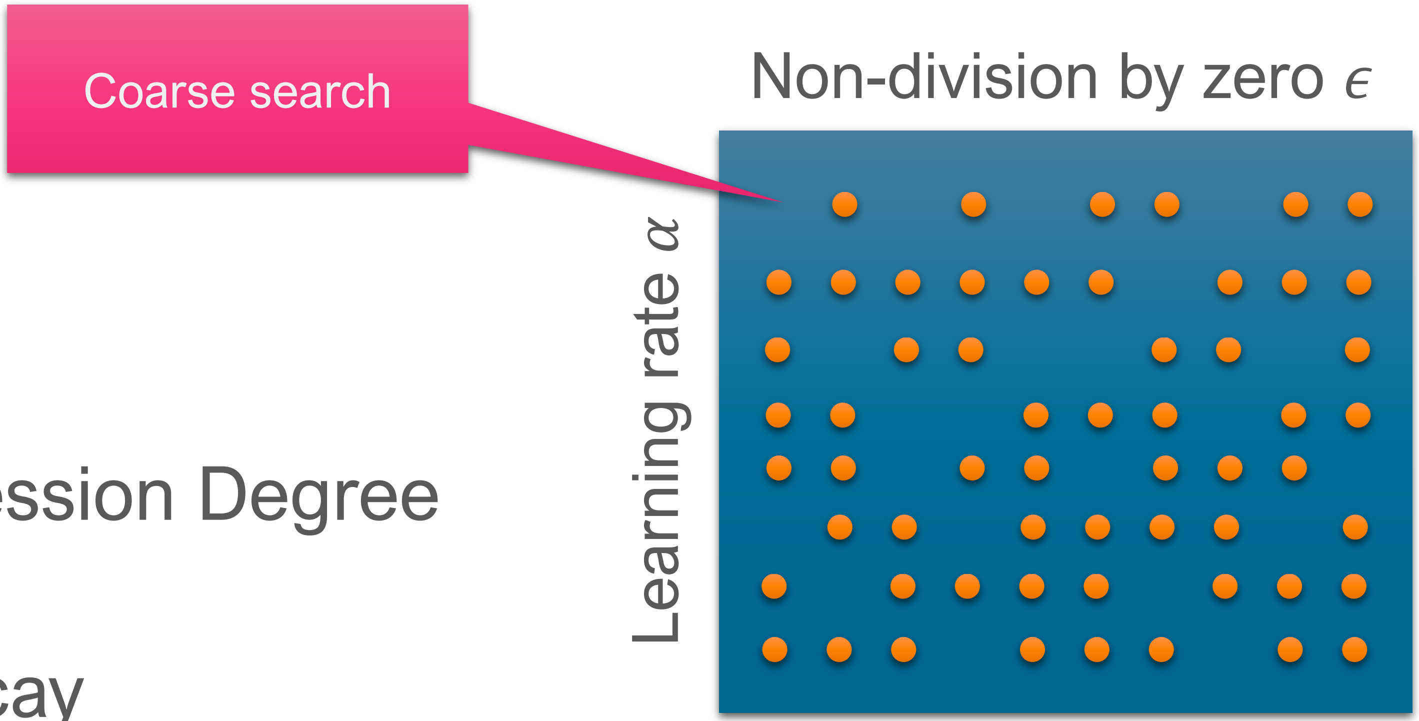
A better approach



Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

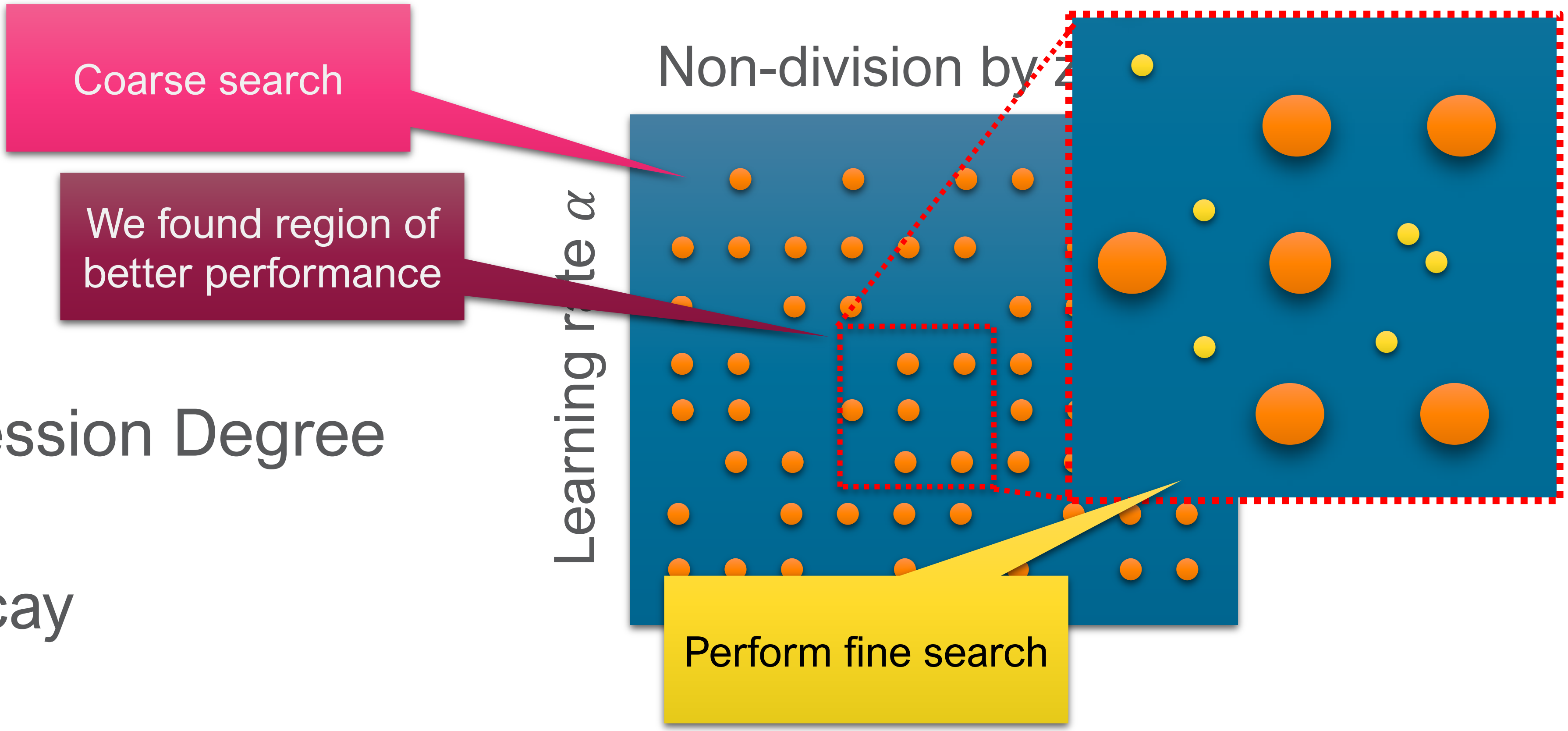
An even better approach



Hyperparameters

- Learning rate α
- Mini-batch size
- Decision Trees
  - Tree depth
  - Bagging Yes/No
  - Size of Forest
- Polynomial Regression Degree
- Regularization λ
- Learning rate decay

An even better approach

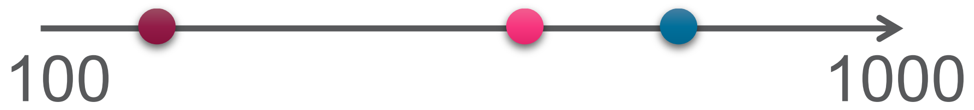


Scaling Hyperparameter Search Space

- Number of trees in a RandomForest T

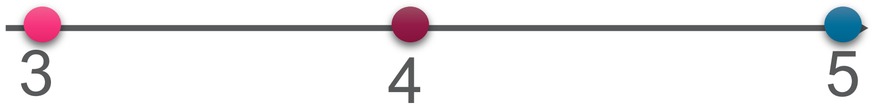
Random samples are ok.

$$T \in \{100, 200, \ldots, 1000\}$$



- Tree Depth L

$$L \in \{3 \ \ldots, 5\}$$

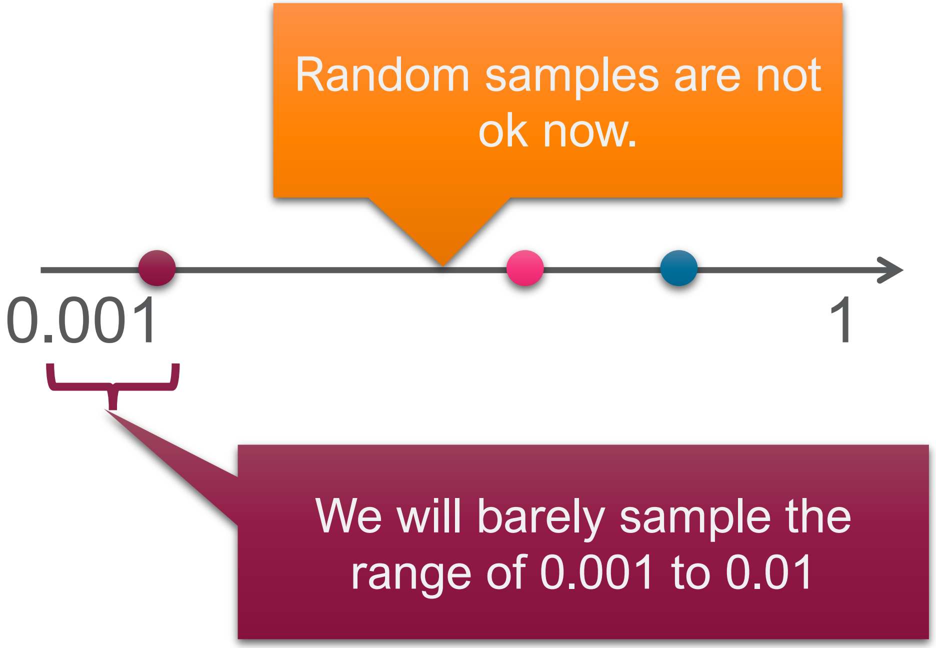


Scaling Hyperparameter Search Space

- Learning rate α

$$\alpha \in \{0.001, \dots, 1\}$$

- $\alpha = np.rand\left(range(0.001, 1.0)\right)$



Scaling Hyperparameter Search Space

- Learning rate α

$$\alpha \in \{0.001, \dots, 1\}$$

- $\alpha = np.rand\left(range(0.001,1.0)\right)$

Log of ending point

$\log(1)$

Log of starting point

$\log(0.001)$

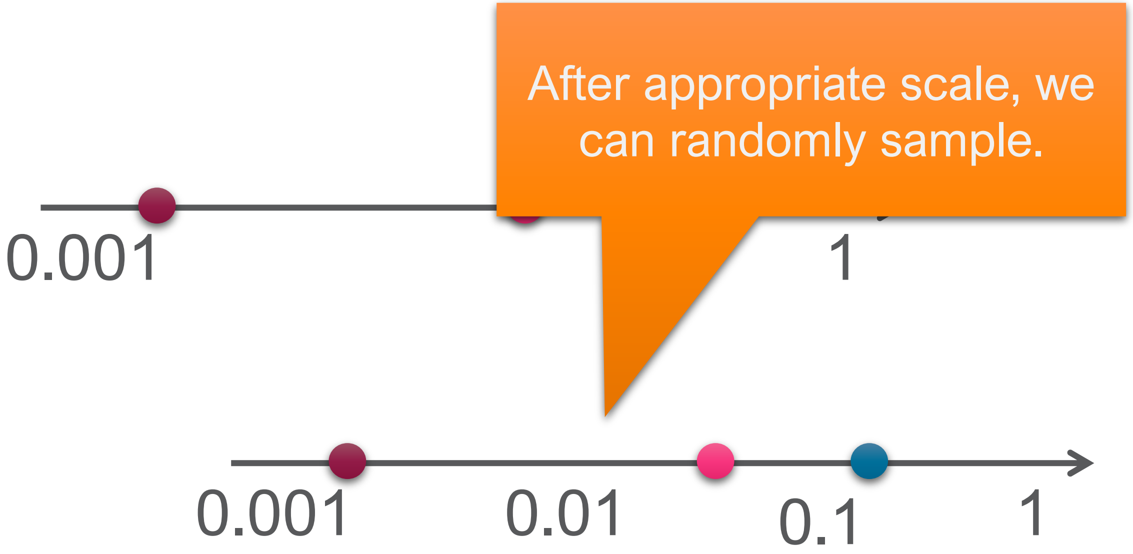
- Instead do $\alpha = np.power\left(10, np.rand\left(range(-3,0)\right)\right)$

Scaling Hyperparameter Search Space

- Learning rate α

$$\alpha \in \{0.001, \dots, 1\}$$

- $\alpha = np.rand(range(0.001{,}1.0))$



- Instead do $\alpha = np.power(10, np.rand(range(-3{,}0)))$

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Pop Quiz

Which of the following is a model parameter? (Select all that apply)

A. Decision Tree Split Features and Thresholds

B. Tree Depth

C. Logistic regression hypothesis coefficients

D. Learning rate

Notebook Time

Review

- Basic Data Wrangling Steps
  - Missing values
  - Scaling
  - Encoding of categorical values
- Pipelines
- Hyperparameter tuning
  - GridSearch
  - RandomSearch

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Next Lecture

- Feature Selection
- Model Explainability

Exam 1 Review

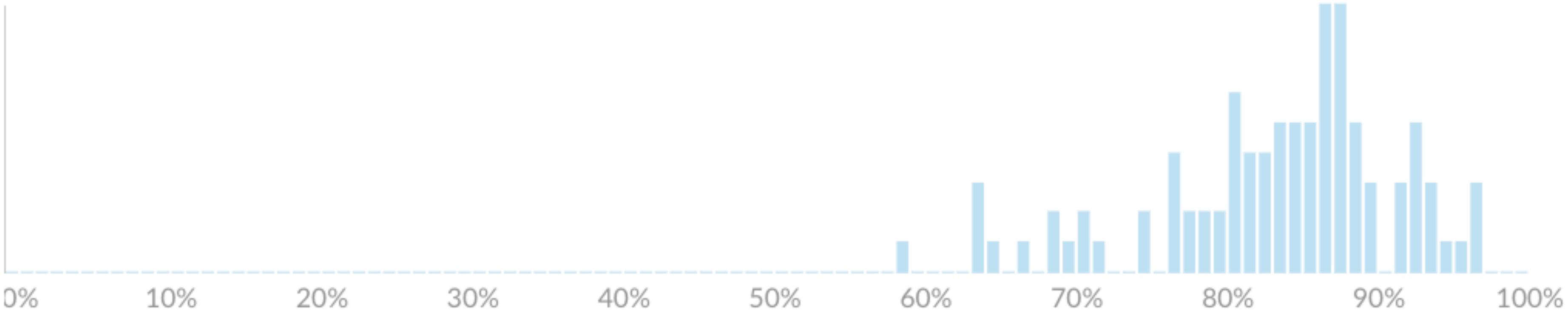
THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Overall Exam Statistics

Quiz Summary

Section Filter ▾ | Student Analysis | Item Analysis

Average Score	High Score	Low Score	Standard Deviation	Average Time
84%	97%	59%	2.44	01:01:97



Question 30 — 0.75 / 1 pts

Which of the following techniques can be used to prevent overfitting in decision trees?

- Correct! — [x] Pruning the tree
- [ ] Decreasing the learning rate
- Correct Answer — [ ] Using GainRatio instead of Information Gain.
- Correct! — [x] Increase minimum number of samples per leaf node.
- Correct! — [x] Decrease tree depth.
- [ ] Increasing the tree depth

Additional Comments:

Gain Ratio

- Addresses wide trees and helps with overfitting
- Penalizes node splits for features with several categories
  - E.g., Date column
- When the number of child nodes is 10x, SplitInfo is 2x

$$GainRatio(\mathcal{D}, V) = \frac{Gain(\mathcal{D}, V)}{SplitInfo(\mathcal{D}, V)}$$

$$SplitInfo(\mathcal{D}, V) = -\sum_{v \in V} \frac{|\mathcal{D}_v|}{|\mathcal{D}|} \log_2\left(\frac{|\mathcal{D}_v|}{|\mathcal{D}|}\right)$$

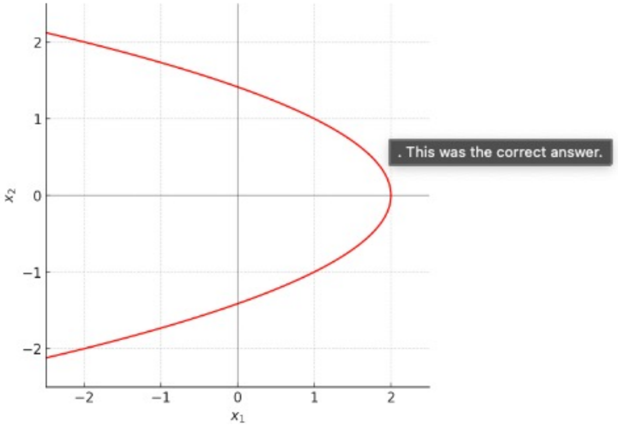
Question 29

0 / 1 pts

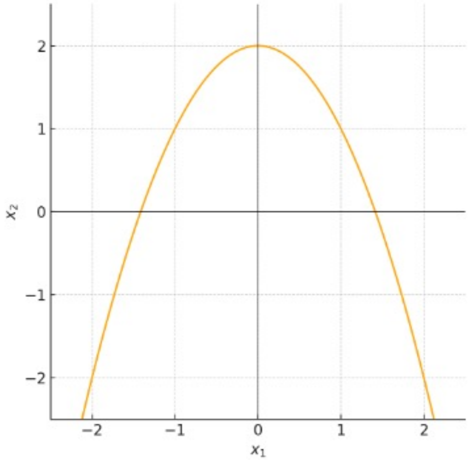
What is the decision boundary of a logistic regression model defined as

$\sigma(z) = \sigma(x_2^2 + x_1 - 2)$?

Correct Answer



You Answered



Question 28

0 / 1 pts

For linear regression model with basis $\hat{y} = \theta_0 + \theta_1 x_1$, what is the new basis for the model if the only feature is categorical; $x_1 \in \{Black, Blue, Green\}$?

- $\hat{y} = \theta_0 + \theta_1 x_1$: No change in basis.
- Correct Answer: $\hat{y} = \theta_0 + \theta_1 x_{black} + \theta_2 x_{Blue}$: We replaced the original feature vector with two new vectors for the black and blue categories.
- You Answered: $\hat{y} = \theta_0 + \theta_1 x_{black} + \theta_2 x_{Blue} + \theta_3 x_{Green}$: We replaced the original feature vector with three new vectors, one for each category.
- $\hat{y} = \theta_1 x_{black} + \theta_2 x_{Blue} + \theta_3 x_{Green}$: We replaced the original feature vector with three new vectors, one for each category, and removed the intercept parameter.

Additional Comments:

https://ttpoll.com/p/618878

Match each data split concept to its purpose.

You Answered	Training Set	Used to assess variance and a
		Used to assess bias, and adjust model parameters.
You Answered	Validation Set	Used for final hyperparameter
		Used to assess variance and adjust hyperparameters.
Correct!	Test Set	Used for final evaluation of th

Consider a logistic regression model with L1 regularization (Lasso). If the penalty parameter $\lambda = 2.0$ and the parameter vector is $\theta = [1.0, -1.0, 2.0]$, calculate the L1 regularization term $R(W)$.

Where the cost is $J(W) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y, \hat{y}) + \frac{\lambda}{m}R(W)$

You Answered

8

4 (with margin: 0)

Consider a logistic regression model with L2 regularization (Ridge). If the penalty parameter $\lambda = 2.0$ and the parameter vector is $\theta = [1.0, -1.0, 2.0]$, calculate the L2 regularization term $R(W)$. =6

Where the cost is $J(W) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y, \hat{y}) + \frac{\lambda}{m} R(W)$

Given the input matrix X with n samples and m features and the target vector y below.

$$X = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix}, \; y = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

What is the value of sample 3 feature 1? Assumes sample index $i \in [1, n]$.

Correct!

9

9 (with margin: 0)

Consider two models. Model A has high bias and low variance, while Model B has low bias and high variance. If the error on the training set for Model A is 10% and for Model B is 5%, which model would likely generalize better to unseen data?

Correct Answer — ○ Model A

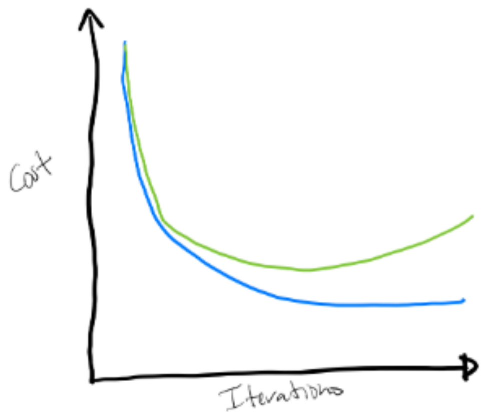
You Answered — ◉ Model B

Which of the following strategies is more likely to reduce irreducible error?

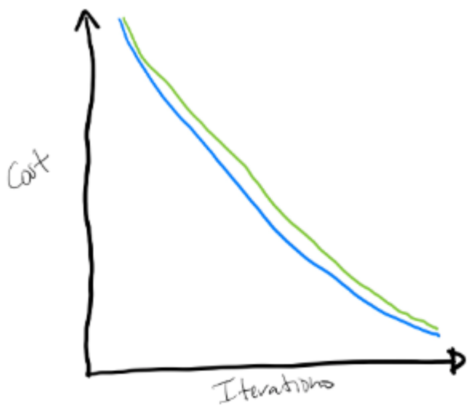
- Fine-tuning the model's hyperparameters
- **Correct Answer:** Reducing noise in the data collection process
- Increasing the amount of training data
- **You Answered:** Using a more complex model

We trained four regularized logistic regression models. Below, we show each model's training and validation sets cost vs. iterations plots. Select the models requiring higher regularization.

Correct!



You Answered



If we have a linear machine learning problem with true hypothesis `f(x)` and `f(x)` is known. What technique guarantees the lowest error between the targets y and corresponding model predictions $\hat{y}$?

You Answered

- Linear Regression

Correct Answer

- Bayes Optimal Classifier
- Decision Trees
- Logistic Regression

In a binary classification problem, which of the following scenarios would result in the highest information gain when splitting on a feature?

You Answered

- [x] One child node contains all the samples, and the other is empty

  Same Impurity as parent.

You Answered

- [x] Both child nodes have approximately equal proportions of each class

  Close to max impurity for both children.

- [ ] Both child nodes have equal numbers of each class

Correct Answer

- [ ] One child node is pure (contains only one class), and the other is evenly split between classes

Which of the following are examples of data leakage?

- [ ] Problem: Insulin dose prediction.

  Sample: Last hour 10-minute moving average (i.e., window of values) of glucose levels.

  Data Split: Separate dataset between past and future. Leave future samples for validation and test sets. Use past samples for training.

Correct!

- [x] Problem: Facial recognition

  Sample: Facial image (Note: Multiple samples per participant)

  Data split: Randomly assign samples to training, validation, and test sets.

  . You selected this answer. This was the correct answer.

- [ ] Problem: House Market Price Prediction.

  Sample: House specifications and sale price. No time data. One sample per house.

  Data Split: Randomly assign samples to training, validation, and test sets.

Correct!

- [x] Problem: Stock market change prediction.

  Sample: Five-day stock price values (i.e., window of values).

  Data Split: Randomly assign samples to training, validation, and test sets.

In polynomial regression, we enhance the input matrix X by adding nonlinear features. For an input matrix $X = [1, 2, 3]$, what is the enhanced input matrix if we apply a polynomial transformation of degree $d = 2$. Recall that the first column corresponds to the intercept parameter.

- [1, 2, 3, 6]
- [1, 2, 3]
- You Answered: [1,2,3,4,9]
- Correct Answer: [1,2,3,4,6,9]

Which of the following indicates we should stop growing the tree at a particular node?

Correct Answer

- [ ] A statistical test determines that a split distribution is the same as the parent distribution.
- [ ] The node has non-zero Gini or Entropy value.

Correct!

- [x] Features values are the same for all samples.

Correct!

- [x] The node has Gini or Entropy value equal to zero.

When assessing training model error, the [loss] function is the average of the [cost] function over the entire training dataset.

Answer 1:

You Answered: loss

Correct Answer: cost

Answer 2:

You Answered: cost

Correct Answer: loss

Question

1 pts

Which of the following statements about linear regression are true?

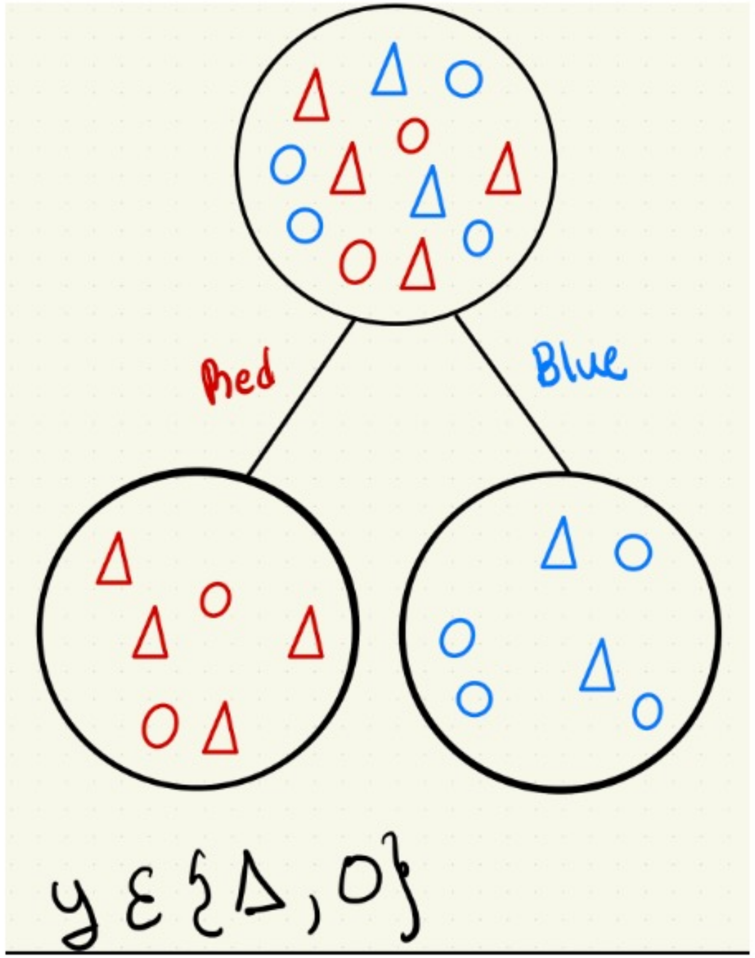
- Correct Answer — It is good classifier for balanced binary classification problems.
- Correct Answer — It is not ideal for classification because its output is unbounded.
- It is ideal for classification because it provides a measurable distance between nominal class categories.
- Correct Answer — Its predictions are easier to interpret.

move/copy question to another bank

If the impurity of a parent node is $I_H(D_p) = 0.76$, and we found a feature and threshold that splits the parent dataset in a left child node with samples from a single class and a right child node with an equal number of samples per class. The probability of the samples landing in the left and right nodes is 0.4 and 0.6, respectively. What is the information gain for this split?

$$IG(D_p, V) = I(D_p) - \frac{N_{Left}}{N_p} I(D_{Left}) - \frac{N_{Right}}{N_p} I(D_{Right})$$ =0.16

For the following section of a tree:

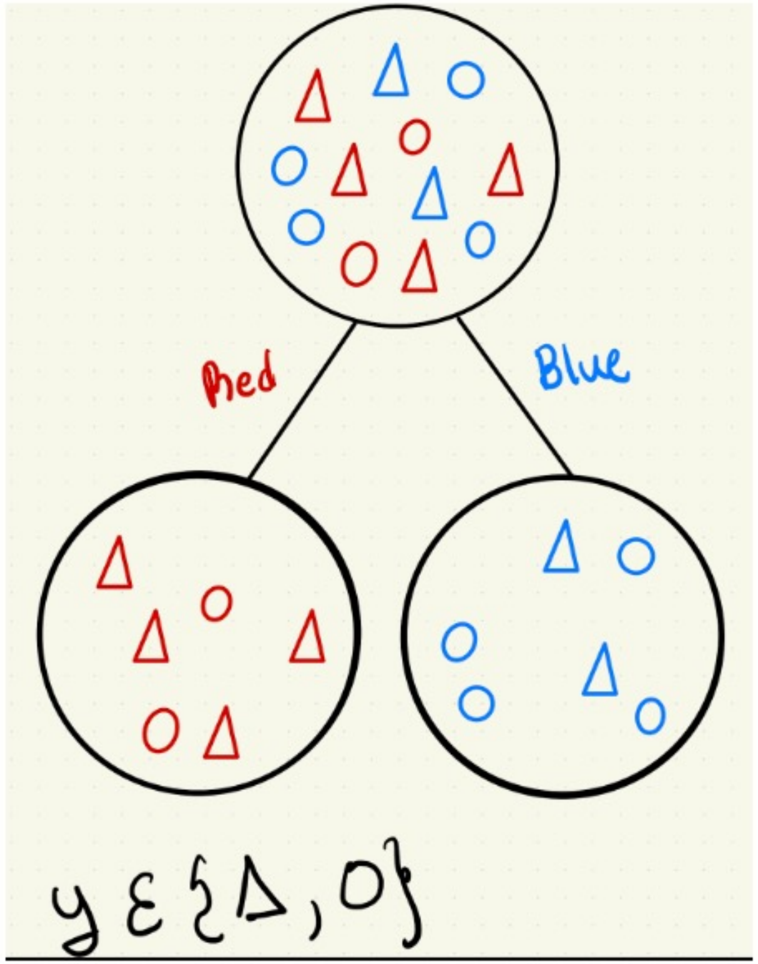


$$-\frac{4}{6}\log_2\left(\frac{4}{6}\right)-\left(1-\frac{4}{6}\right)\log_2\left(1-\frac{4}{6}\right)=0.91829$$

What is the Entropy of the left child node? =0.92

$$I_H = -p\log_2(p) - (1-p)\log_2(1-p)$$

For the following section of a tree:



=0.92

What is the Conditional Entropy of the child nodes?

$$I_H = -p\log_2(p) - (1-p)\log_2(1-p)$$

Entropy of left or right child:

$$-\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \left(1-\frac{4}{6}\right)\log_2\left(1-\frac{4}{6}\right) = 0.91829$$

$$I_H(D|V) = \sum_{v \in V} p(V=v) I_H(D|V=v)$$

$$V \in \{Red, Blue\}$$

Given the input matrix X with n samples and m features and the target vector y below.

$$X = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix}, \; y = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

What are the features values for sample 3. Assumes sample index $i \in [1, n]$.

- $x_3^T = [9, 10, 11, 12]$
- $x_3^T = [3, 7, 11]$
- $x^{(3)T} = [3, 7, 11]$
- $x^{(3)T} = [9, 10, 11, 12]$

Helper Slides