



Lecture 2

BU.330.775 Machine Learning

Minghong Xu, PhD.
Associate Professor

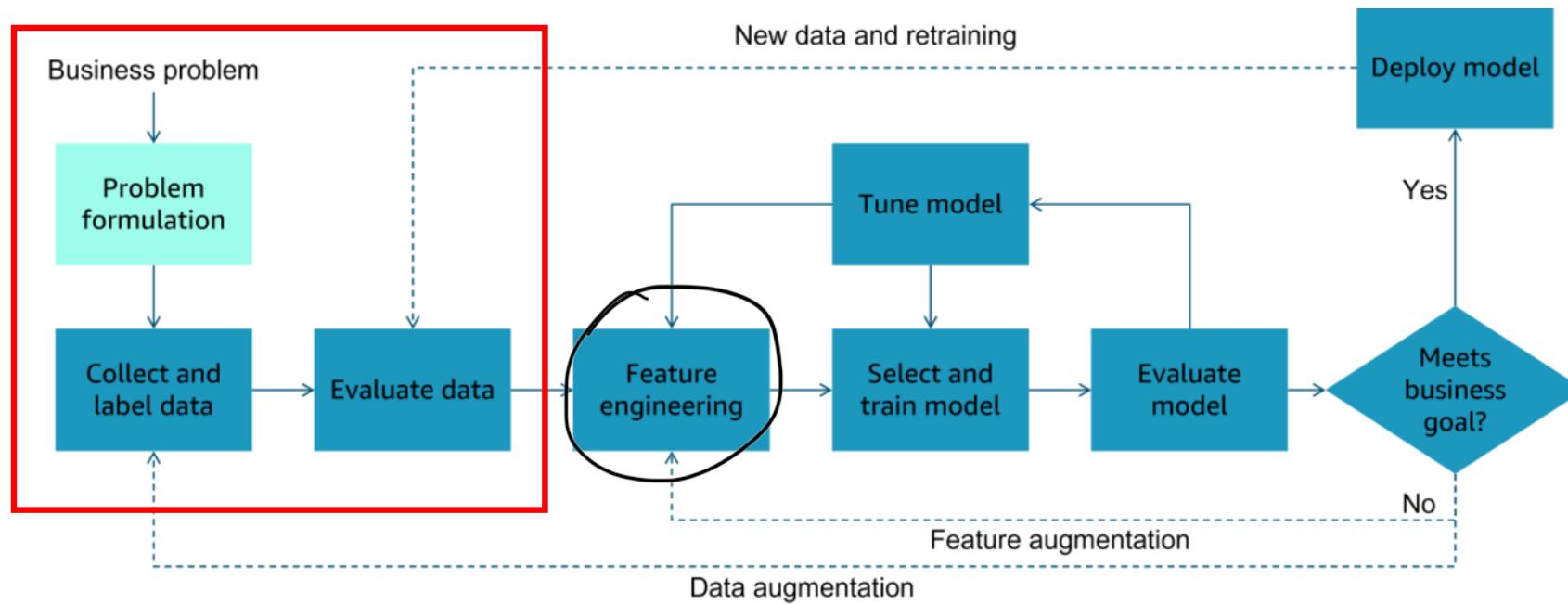
Today's Agenda



- » Data preparation and preprocessing
- » Exploratory data analysis
- » Hands-on practice on California housing data

- » Reminder: homework 1 due before class

Machine Learning Pipeline



Problem Formulation



- » *Business objective: how to use? benefit?* ✓
- » *Is it machine learning problem?* ✗
- » *Supervised? Unsupervised?*
- » *Target to predict?*
- » *Data?*
- » *Current solution?*



Example: Credit Card Fraud

- » Business objective:
- » Machine learning:
- » Supervised/unsupervised:
- » Target:
- » Data:



Data Preparation

Data Terminologies



- » Observation: instance/example/data point
 - A single record that represents one example of the problem being studied
- » Feature: each piece of information included in the data representation
 - E.g. age, weight, height, diet, exercise or not...
 - Continuous feature
 - Categorical/discrete feature, not numerical
- » Target: output variable

Questions to Consider



- » *How much data do you have?*
- » *Where is it?*
- » *Do you have access?*
- » *How to bring all data into one centralized repository?*
- » *Is your data representative?*

Popular Open Data Repositories



- » OpenML: <https://openml.org>
- » Kaggle: <https://www.kaggle.com>
- » PapersWithCode: <https://paperswithcode.com/datasets>
- » UC Irvine: <https://archive.ics.uci.edu>
- » AWS: <https://registry.opendata.aws>
- » TensorFlow: <https://www.tensorflow.org/datasets>
- » Nasdaq: <https://data.nasdaq.com/institutional-investors>
- » World Health Organization: <https://www.who.int/data/collections>
- » U.S. Census Bureau: <https://www.census.gov/data/datasets.html>

good
resources

Some Data Portals



- » <http://dataportals.org>
- » https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research
- » Social media posts:
 - » <https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public>
 - » <https://www.reddit.com/r/datasets/>

Pipeline



- » *How much data do you have?*
- » *Where is it?*
- » *Access?*
- » ***How to bring all data into one centralized repository?***

- » Data pipeline: a sequence of data processing components/steps
- » A popular one: Extract, Transform, Load (ETL)

Data Evaluation



- » *How much data do you have?*
- » *Where is it?*
- » *Access?*
- » *How to bring all data into one centralized repository?*
- » ***Is your data representative? Good quality?***

- » **Data evaluation!**

Evaluation Tasks



» Data format

- Common file formats for machine learning: csv, ~~text~~, json, etc. image, audio , video

» Data types → Numeric, Categorical, Text,

- Structured, unstructured Time series, Image
audio, video.

JPEG MP3 MD4
PNG WAV AVI
MKV

» Descriptive statistics

- Statistical: max, min, mean, median, std, var, etc.
- Correlations

» Data visualization



Data Preprocessing

Training and Testing



- » Most machine learning models have two stages: training and testing
- » Training data is used to train the model, to learn patterns
- » Testing data is used to evaluate the model's performance after training
- » Training and testing sampling: a process that splits a data set into two parts, a training set and a testing set

Validation



» Similar to preparing an exam

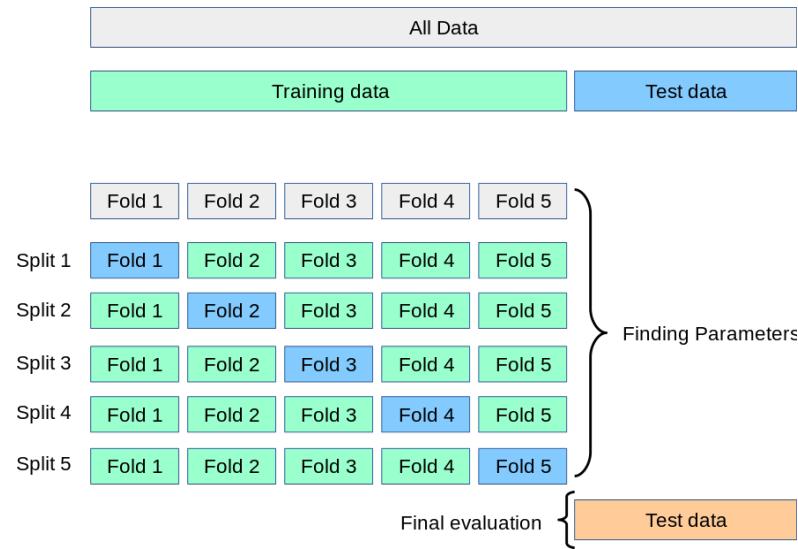
- Do lots of exercises => Training
- Mock exam => Validation
- Final exam => Test



Cross-Validation



- » Machine learning models need sufficient training, thus sufficient usage of data
- » Cross-validation is useful for small datasets



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



Sampling Strategy 1: Stratifying

» Divide subjects into subgroups called strata based on characteristics that they share

- Race
- Gender
- Educational attainment
- Etc.

Notes ↗

» Once divided, each subgroup is randomly sampled

» Also called Stratified Sampling to ensure each group (strata) has representation in Sampling in equal proportion in test and train data set



Sampling Strategy 2: Shuffling

- » Process of randomly rearranging the order of the rows/instances in a dataset
- » Important especially when splitting data into training and testing sets or cross-validation
- » Help to ensure that the model trained is not biased by the order of the data
- » When not to shuffle: time-series



Handling Categorical Data

- » Most machine learning algorithms prefer to work with numbers
- » **Label encoding:** each category into a unique integer
 - When: if categorical variable has a natural order, e.g. low, medium, high
 - Also called ordinal encoding
- » **One-hot encoding:** convert into a set of binary (0/1) columns, one for each category
 - When: there is no inherent order



Handling Missing Value

- » Most machine learning algorithms can not work with missing features
 - Need to take care before you can run the model
- » **Three options:**
 - » Get rid of the data points with missing values
 - » Get rid of the attributes with missing values
 - » **Imputation:** set the missing values to some value, e.g. zero, mean, median...

Examples



» Transaction or account activity analysis

- Zero can be used to indicate no activity/transaction/balance

» Temperatures, blood pressure readings, other (usually) normal distributed cases

- Mean can be used to represent average case

Note

» House price prediction, outliers are significant or distribution is skewed

- Median can be used to represent typical values

Simple Impute (Strategy = median)

» *What are the potential issues for simply dropping?*

Handling Outliers



» Outliers may skew the models, longer training, less accurate

» Identifying outliers can be a difficult task

- Statistical methods: Z-scores, quartiles
- Visualizations: box plot, scatter plot, histogram
- Machine learning: clustering, isolation forest

» Once identified, drop or keep?

» General “drop” cases:

- Data errors
- Sensitive models such as regressions

Notes {

273 consider as outlier

$$IQR : Q_3 - Q_1$$

$$\text{outlier} < Q_1 - 1.5 \cdot IQR$$

$$\text{outlier} > Q_3 + 1.5 \cdot IQR$$

① k-Mean clustering

② DBSCAN

Feature Scaling



- » Most machine learning algorithms do not perform well when attributes have very different scales
 - Results may be dominated by the large scale
- » Min-max scaling/normalization: transform into a given range from 0 to 1
for data that doesn't follow Gaussian distribution
- » Standardization: transform into standard normal distribution
 - Not a specific range, less affected by outliers
- » It is important to apply exactly the same transformation to the training set and the test set!



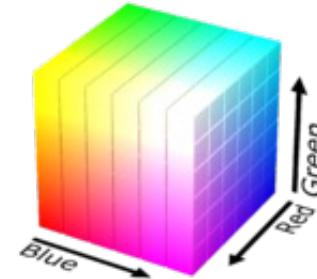
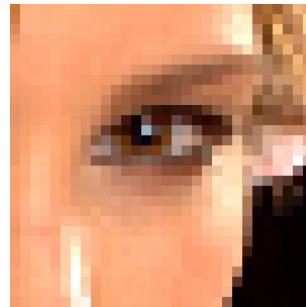
Unstructured Data: Images

Image Data



- » *What is an image?*
- » A bunch of numbers between 0 to 255
- » A 3-dimensional array

$(0, 0, 0)$ black
 $(255, 255, 255)$ white



- » <https://convertacolor.com>

Grayscale Image



➤ A grayscale (or graylevel) image: **the only colors are shades of gray**

- One number for each pixel
 - Black/white image

| | | | | | | | | | | | | | | | | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 206 | 206 | 245 | 244 | 253 | 247 | 245 | 136 | 151 | 255 | 255 | 255 | 255 | 255 | 236 | 209 | 231 | 254 | 254 | 255 | 254 | 255 | 255 | 254 | 247 | |
| 143 | 160 | 234 | 254 | 255 | 254 | 255 | 118 | 103 | 229 | 255 | 135 | 232 | 193 | 74 | 52 | 68 | 173 | 255 | 254 | 255 | 255 | 255 | 254 | 234 | |
| 192 | 154 | 75 | 200 | 245 | 255 | 255 | 116 | 90 | 84 | 81 | 35 | 91 | 53 | 45 | 43 | 56 | 141 | 213 | 253 | 255 | 255 | 255 | 254 | 187 | |
| 109 | 96 | 143 | 223 | 255 | 255 | 255 | 127 | 75 | 41 | 35 | 31 | 24 | 36 | 45 | 44 | 44 | 81 | 174 | 234 | 252 | 254 | 249 | 231 | 253 | |
| 67 | 107 | 196 | 236 | 255 | 255 | 255 | 105 | 34 | 54 | 24 | 33 | 32 | 30 | 34 | 33 | 53 | 105 | 140 | 231 | 247 | 249 | 255 | 255 | 253 | |
| 52 | 52 | 143 | 219 | 255 | 255 | 253 | 112 | 26 | 33 | 24 | 46 | 75 | 72 | 73 | 71 | 58 | 57 | 87 | 90 | 136 | 228 | 186 | 253 | 247 | 255 |
| 78 | 56 | 56 | 234 | 255 | 255 | 255 | 118 | 11 | 27 | 74 | 99 | 101 | 160 | 162 | 172 | 173 | 172 | 172 | 173 | 172 | 172 | 173 | 172 | 97 | |
| 38 | 43 | 48 | 52 | 147 | 255 | 259 | 58 | 41 | 81 | 129 | 145 | 160 | 160 | 173 | 178 | 179 | 179 | 179 | 177 | 177 | 177 | 172 | 111 | 32 | |
| 40 | 43 | 37 | 30 | 245 | 272 | 33 | 65 | 110 | 139 | 143 | 162 | 171 | 171 | 179 | 182 | 182 | 187 | 183 | 173 | 172 | 164 | 73 | 46 | 255 | |
| 34 | 41 | 51 | 69 | 250 | 251 | 76 | 70 | 129 | 143 | 152 | 163 | 171 | 171 | 177 | 182 | 190 | 194 | 189 | 170 | 170 | 151 | 137 | 265 | 255 | |
| 45 | 51 | 65 | 116 | 237 | 181 | 154 | 161 | 180 | 144 | 140 | 154 | 176 | 178 | 177 | 173 | 185 | 185 | 185 | 184 | 180 | 146 | 143 | 254 | 249 | |
| 36 | 54 | 72 | 74 | 171 | 188 | 156 | 63 | 131 | 134 | 134 | 155 | 160 | 161 | 173 | 179 | 189 | 193 | 190 | 185 | 187 | 182 | 193 | 146 | 253 | |
| 32 | 38 | 54 | 54 | 159 | 250 | 126 | 57 | 129 | 138 | 140 | 151 | 166 | 168 | 171 | 171 | 181 | 188 | 186 | 186 | 183 | 183 | 180 | 12 | 342 | |
| 32 | 32 | 72 | 129 | 212 | 254 | 116 | 65 | 121 | 104 | 104 | 93 | 134 | 136 | 170 | 162 | 162 | 121 | 143 | 145 | 145 | 253 | 253 | 253 | 255 | |
| 62 | 62 | 116 | 107 | 179 | 241 | 73 | 60 | 102 | 92 | 111 | 109 | 103 | 94 | 147 | 141 | 127 | 128 | 135 | 147 | 142 | 200 | 90 | 122 | 207 | 217 |
| 144 | 178 | 261 | 230 | 212 | 70 | 67 | 115 | 86 | 78 | 62 | 85 | 88 | 139 | 192 | 191 | 153 | 83 | 99 | 141 | 164 | 201 | 79 | 192 | 245 | 248 |
| 127 | 145 | 150 | 204 | 213 | 197 | 95 | 133 | 122 | 133 | 136 | 136 | 136 | 136 | 136 | 136 | 136 | 136 | 136 | 136 | 137 | 147 | 147 | 147 | 147 | |
| 87 | 112 | 100 | 79 | 85 | 82 | 65 | 72 | 145 | 151 | 153 | 138 | 129 | 140 | 191 | 190 | 193 | 175 | 174 | 192 | 198 | 208 | 127 | 163 | 231 | 219 |
| 83 | 83 | 104 | 139 | 128 | 78 | 78 | 132 | 146 | 159 | 159 | 111 | 164 | 164 | 205 | 198 | 191 | 195 | 201 | 202 | 202 | 203 | 253 | 245 | 249 | 234 |
| 78 | 78 | 73 | 97 | 74 | 73 | 106 | 127 | 140 | 152 | 155 | 125 | 97 | 150 | 156 | 174 | 174 | 183 | 196 | 203 | 202 | 168 | 246 | 254 | 255 | 254 |
| 72 | 45 | 83 | 59 | 46 | 52 | 48 | 74 | 73 | 137 | 146 | 149 | 132 | 78 | 70 | 34 | 141 | 168 | 169 | 207 | 204 | 216 | 234 | 246 | 222 | 243 |
| 22 | 69 | 69 | 59 | 89 | 46 | 47 | 137 | 147 | 144 | 161 | 144 | 140 | 150 | 156 | 157 | 183 | 182 | 196 | 201 | 205 | 214 | 194 | 178 | 185 | 183 |
| 45 | 49 | 77 | 89 | 50 | 89 | 43 | 61 | 109 | 127 | 142 | 148 | 113 | 120 | 121 | 145 | 149 | 148 | 169 | 172 | 181 | 202 | 203 | 204 | 202 | 174 |
| 76 | 72 | 79 | 74 | 59 | 58 | 47 | 43 | 70 | 121 | 132 | 116 | 89 | 111 | 146 | 143 | 122 | 149 | 120 | 195 | 197 | 195 | 198 | 183 | 184 | 184 |
| 107 | 121 | 123 | 105 | 79 | 79 | 36 | 111 | 122 | 130 | 114 | 157 | 145 | 170 | 163 | 101 | 120 | 170 | 207 | 187 | 197 | 146 | 145 | 152 | 155 | 158 |
| 117 | 134 | 123 | 135 | 105 | 21 | 21 | 38 | 88 | 115 | 121 | 135 | 128 | 141 | 142 | 166 | 202 | 212 | 215 | 210 | 195 | 177 | 152 | 133 | 59 | |
| 191 | 118 | 116 | 128 | 122 | 111 | 29 | 28 | 56 | 100 | 131 | 130 | 141 | 151 | 159 | 181 | 180 | 205 | 192 | 197 | 196 | 194 | 147 | 143 | 141 | 144 |
| 117 | 120 | 130 | 130 | 130 | 18 | 30 | 44 | 58 | 70 | 102 | 135 | 147 | 168 | 196 | 212 | 215 | 210 | 195 | 177 | 152 | 133 | 59 | 58 | 126 | 151 |
| 155 | 123 | 137 | 145 | 101 | 27 | 54 | 58 | 45 | 75 | 103 | 135 | 175 | 178 | 190 | 216 | 208 | 169 | 131 | 116 | 144 | 203 | 74 | 6 | 121 | 149 |
| 108 | 108 | 124 | 132 | 105 | 44 | 31 | 55 | 47 | 54 | 58 | 101 | 147 | 144 | 136 | 145 | 140 | 195 | 146 | 187 | 196 | 185 | 169 | 143 | 144 | 146 |
| 97 | 97 | 96 | 104 | 76 | 74 | 33 | 31 | 49 | 41 | 49 | 58 | 74 | 53 | 66 | 69 | 89 | 150 | 189 | 192 | 198 | 103 | 168 | 106 | 131 | 131 |
| 103 | 102 | 97 | 73 | 39 | 35 | 23 | 42 | 50 | 62 | 46 | 92 | 50 | 51 | 57 | 52 | 123 | 157 | 185 | 205 | 169 | 162 | 93 | 16 | 105 | 101 |

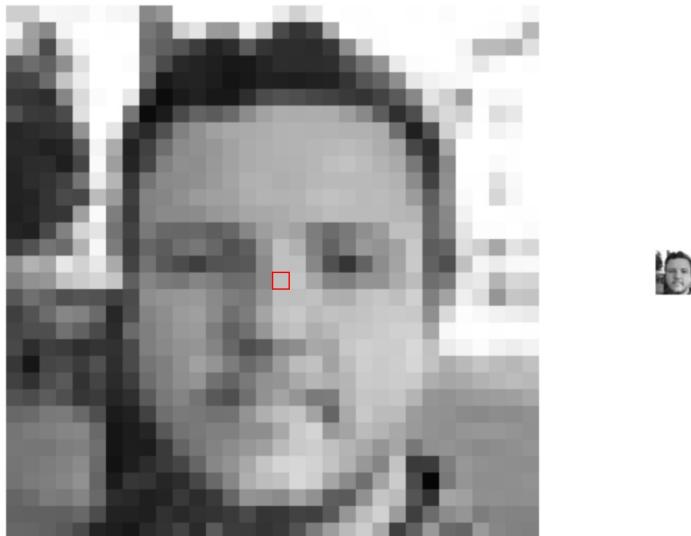


Image Representations



» <https://www.boxentriq.com/code-breaking/pixel-values-extractor>

» Size: **width * height * # color channels**

- 3 color channels for colored images
- 1 color channel (grayscale) for black/white images

» That's how many numbers we use to represent an image

» *How can we know the size of an image?*

ImageNet Dataset



- » <https://www.image-net.org>
- » More than 14 million images, 21000 categories
- » Hand-annotated
- » Initiated by AI Researcher, Dr. Fei-Fei Li, in 2006
 - Expand and improve the data to train AI models
- » Amazon Mechanical Turk to help with classification
 - Average: 50 images per minute





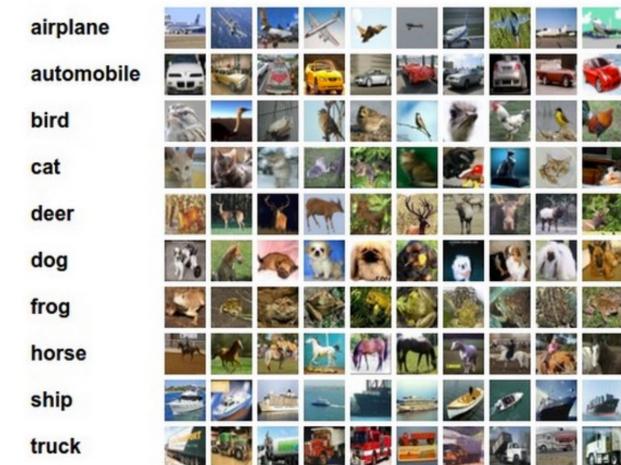
MNIST Dataset

- » Modified National Institute of Standards and Technology Database
 - » A large database of handwritten digits
 - Commonly used for training various image processing systems
 - Taken from American high school students
 - Black/white images (grayscale levels), 28x28 pixel
 - 60,000 training images and 10,000 testing images

CIFAR-10 Dataset



- » Canadian Institute For Advanced Research, 10 classes
 - Airplane, car, bird, cat, deer, dog, frog, horse, ship, truck
- » Commonly used to train machine learning and computer vision algorithms
- » 60000 images, 6000 images per class
- » Images are 32x32x3 in size
 - width*height*3 color channels



Lab 2: California Housing data



» Pandas package

- Derived from “Panel Data”
- Handle and analyze structured data, such as tabular data, time-series, etc.

» Sklearn package

- Originally scikit-learn, toolkit for scientific computing focus on machine learning
- Library for machine learning
- Built on top of NumPy, SciPy and matplotlib

» **Note that** we follow lecture notes sequence in this exercise, the actual preprocessing sequence in your own project might be different

Python Indentation



- » To nest the code blocks
 - Function definition
 - Conditional statements
 - Loops
- » Use “tab” not multiple “space”
- » Auto controlled in Google Colab
- » Sometimes you need to manually remove/control it

Next Week



- » Supervised machine learning models – part I
- » Model training, regularization, and evaluation
- » Homework 2 due before class

References



» AWS Academy Machine Learning Foundations