Image classification, data-driven approach, knn

ISTD 50.035

Computer Vision

Acknowledgement: Some images are from various sources: UCF, Stanford cs231n, etc.

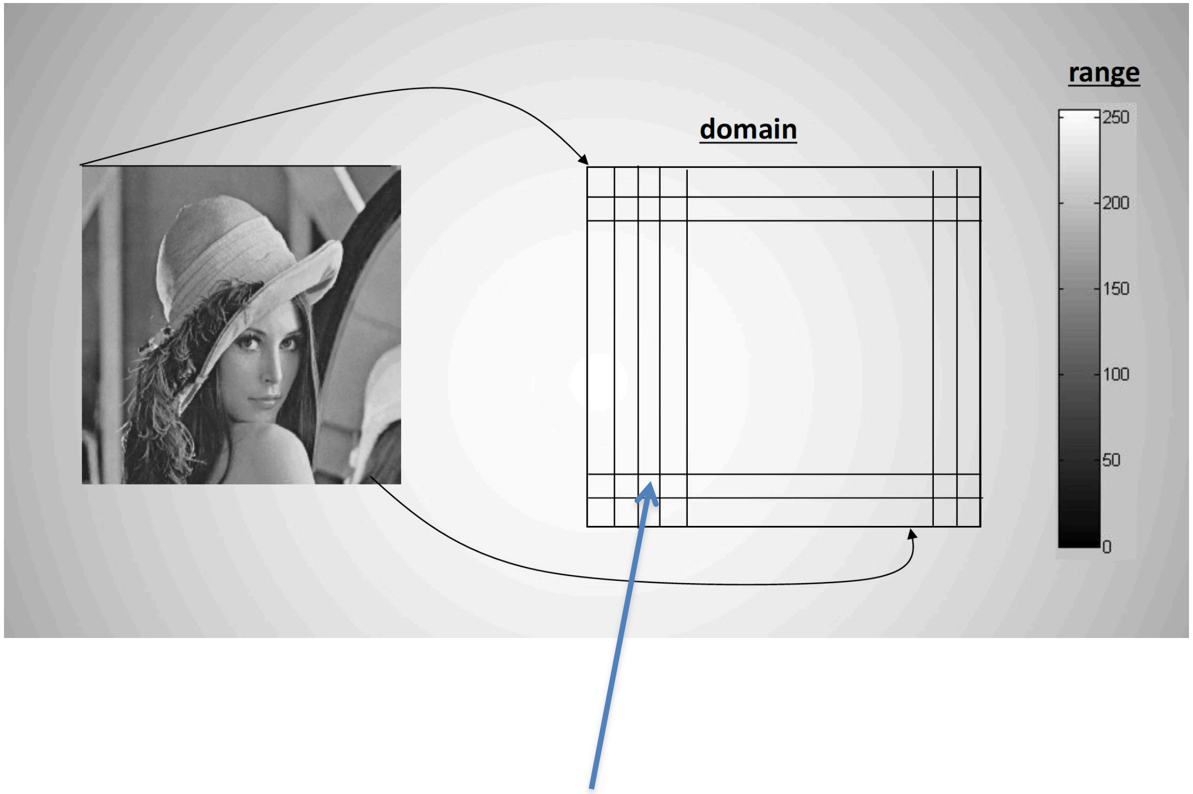
Image is an array of numbers

-Grayscale image

-2D array of numbers (pixels) / matrix

-Number indicates the intensity: [0,255] for 8-bit representation

-Image resolution / number of pixel in an image: 100x100, 1920x1080, etc.



A pixel

0: black, 255: white

Can you recognize this image?

```
[[254 252 248 251 248 243 231 212 188 173 163 159 163 169 162 154]
 [255 253 251 243 227 193 158 145 159 159 154 150 153 158 159 159]
 [246 228 211 170 179 156  70  50  73 138 193 197 165 144 152 173]
 [228 218 186 149 133 130 100  48  47  61 137 192 175 168 170 169]
 [212 200 170 147 121  85 124  65  69 137 185 240 221 174 171 193]
 [222 235 228 192 162 132 150 187 218 217 200 225 228 211 214 214]
 [187 230 244 231 213 236 232 220 222 218 216 213 204 199 207 195]
 [166 173 204 230 245 247 226 193 191 225 218 209 181 132 198 202]
 [189 171 210 210 210 179 210 211 203 192 213 188 106 151 204 184]
 [205 174 219 223 219  99 121 233 231 214 214 212 195 193 151 158]
 [209 137 166 212 204 115 117 202 222 219 210 204 168 114  84 204]
 [214 174 143 155 222 137 139 182 214 193 143 157  58  78 135 198]
 [227 230 222 191 173 129 148 150 184 140 103 147 125 144 165 204]
 [241 239 238 231 216 143 143 163 203 193 175 145 143 164 198 201]
 [250 248 250 245 243 219 135  94 134 156 160 179 194 198 198 198]
 [253 251 249 249 249 246 232 223 229 232 216 213 213 203 198 197]]
```

Can you recognize this image?

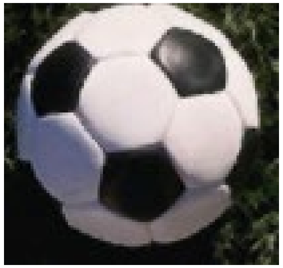
```
[[241 243 237 241 241 241 241 241 241 241 241 241 241 241 241 241]
 [241 244 255 250 241 241 241 241 241 241 241 241 241 241 241 241]
 [248 247 253 248 237 244 241 241 241 241 241 241 241 241 241 241]
 [238 254 255 255 243 236 241 241 241 241 241 241 241 241 241 241]
 [238 255 253 253 245 238 240 250 242 241 241 241 239 236 243 253]
 [237 252 255 236 238 233 239 247 246 241 241 241 246 242 253 248]
 [241 253 243 242 246 241 246 241 237 239 239 242 251 254 241 239]
 [240 241 248 239 239 241 247 240 241 238 237 248 252 252 249 241]
 [240 247 245 238 218 244 240 241 238 240 251 244 253 253 236 241]
 [238 240 247 247 242 249 243 239 241 108 252 247 248 234 241 241]
 [237 233 244 245 249 244 243 242 238 221 245 246 243 241 241 241]
 [238 240 242 249 242 254 255 234 235 233 241 241 241 241 241 241]
 [233 241 244 244 255 250 249 249 236 239 241 241 241 241 241 241]
 [237 222 236 245 246 247 255 235 241 241 241 241 241 241 241 241]
 [244 238 247 243 241 241 234 244 248 241 241 241 241 241 241 241]
 [239 224 241 239 221 230 241 239 241 241 241 241 241 241 241 241]]
```

```
[[243 241 244 243 243 243 243 243 243 243 243 243 243 243 243 243]
 [243 245 175 221 242 243 243 243 243 243 243 243 243 243 243 243]
 [239 163 164 164 194 241 243 243 243 243 243 243 243 243 243 243]
 [234 158 158 145 166 248 243 243 243 243 243 243 243 243 243 243]
 [234 156 142 144 169 215 221 244 243 243 243 243 242 246 242 238]
 [234 161 150 132 182 201 212 223 228 243 243 243 167 168 174 195]
 [226 154 153 197 210 213 218 216 211 210 187 155 156 163 164 243]
 [214 220 219 196 170 204 222 213 219 206 197 162 167 164 242 243]
 [214 216 216 183 187 203 210 215 211 215 208 164 169 161 245 243]
 [212 206 221 205 205 210 205 204 212  72 212 158 169 245 243 243]
 [209 204 216 205 208 197 193 200 201 172 239 244 242 243 243 243]
 [210 213 210 213 195 200 186 183 198 204 243 243 243 243 243 243]
 [194 215 213 202 202 171 155 192 209 241 243 243 243 243 243 243]
 [190 169 195 213 156 171 174 151 213 243 243 243 243 243 243 243]
 [197 196 240 198 213 179 148 203 242 243 243 243 243 243 243 243]
 [198 164 243 248 156 183 243 240 243 243 243 243 243 243 243 243]]
```

```
[[242 244 244 240 242 242 242 242 242 242 242 242 242 242 242 242]
 [242 228 153 222 246 242 242 242 242 242 242 242 242 242 242 242]
 [237 162 154 146 185 242 242 242 242 242 242 242 242 242 242 242]
 [230 146 139 135 143 240 240 242 242 242 242 242 242 242 242 242]
 [222 133 123 118 159 203 207 232 238 242 242 242 248 242 237 236]
 [222 142 130 116 161 176 192 197 211 242 242 242 151 146 162 189]
 [218 135 131 180 182 185 196 193 188 185 179 134 143 154 152 242]
 [201 201 203 176 154 185 200 183 196 185 177 134 148 152 238 242]
 [201 205 200 163 191 185 182 180 174 192 185 145 158 162 242 242]
 [199 195 206 182 175 188 176 174 182  63 195 149 152 247 242 242]
 [193 188 192 174 180 169 170 175 175 149 245 239 242 242 242 242]
 [189 192 185 184 165 179 169 160 171 177 242 242 242 242 242 242]
 [173 196 182 170 188 160 149 168 188 240 242 242 242 242 242 242]
 [167 135 173 180 137 152 153 112 184 242 242 242 242 242 242 242]
 [174 181 238 173 174 161 131 160 233 242 242 242 242 242 242 242]
 [176 145 242 244 122 153 239 242 242 242 242 242 242 242 242 242]]
```

Semantic gap

```
[[242 244 244 240 242 242 242 242 242 242 242 242 242 242 242 242]
 [242 228 153 222 246 242 242 242 242 242 242 242 242 242 242 242]
 [237 162 154 146 185 242 242 242 242 242 242 242 242 242 242 242]
 [230 146 139 135 143 240 240 242 242 242 242 242 242 242 242 242]
 [222 133 123 118 159 203 207 232 238 242 242 242 248 242 237 236]
 [222 142 130 116 161 176 192 197 211 242 242 242 151 146 162 189]
 [218 135 131 180 182 185 196 193 188 185 179 134 143 154 152 242]
 [201 201 203 176 154 185 200 183 196 185 177 134 148 152 238 242]
 [201 205 200 163 191 185 182 180 174 192 185 145 158 162 242 242]
 [199 195 206 182 175 188 176 174 182  63 195 149 152 247 242 242]
 [193 188 192 174 180 169 170 175 175 149 245 239 242 242 242 242]
 [189 192 185 184 165 179 169 160 171 177 242 242 242 242 242 242]
 [173 196 182 170 188 160 149 168 188 240 242 242 242 242 242 242]
 [167 135 173 180 137 152 153 112 184 242 242 242 242 242 242 242]
 [174 181 238 173 174 161 131 160 233 242 242 242 242 242 242 242]
 [176 145 242 244 122 153 239 242 242 242 242 242 242 242 242 242]]
```



The gap between low-level representation of an image (input to an algorithm) and high-level understanding of an image (output)

Image classification

- Given an input image, the algorithm produces one image label from a fixed set of classes (categories)

→ {fish, soccer ball, dog, boat}

- Image recognition (many classes)
  - 1000 categories in IMAGENET Large Scale Visual Recognition Challenge (ILSVRC): zebra, speedboat, lifeboat, …
  - 10,000+ categories in IMAGENET

zebra

speedboat

lifeboat

Image classification

- **Top-n accuracy**
- The algorithm outputs k confidence for each of the k classes

Test image:

Algorithm outputs: {cat, dog, house, mouse} = {0.1, 0.2, 0.0, 0.7}

Top-1 class: {mouse}

Top-2 class: {mouse, dog}

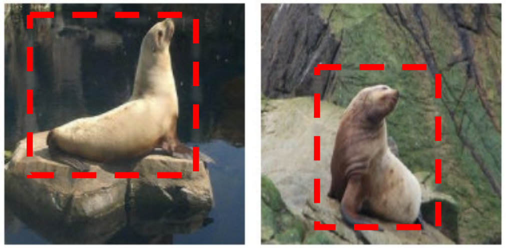
Incorrect for **top-1 accuracy**, correct for **top-2 accuracy** (ground truth is contained in the top-2 class)

- ILSVRC: Top-1, Top-5 accuracy

Number of correct / Number of test image

Image classification is fundamental to many computer vision tasks

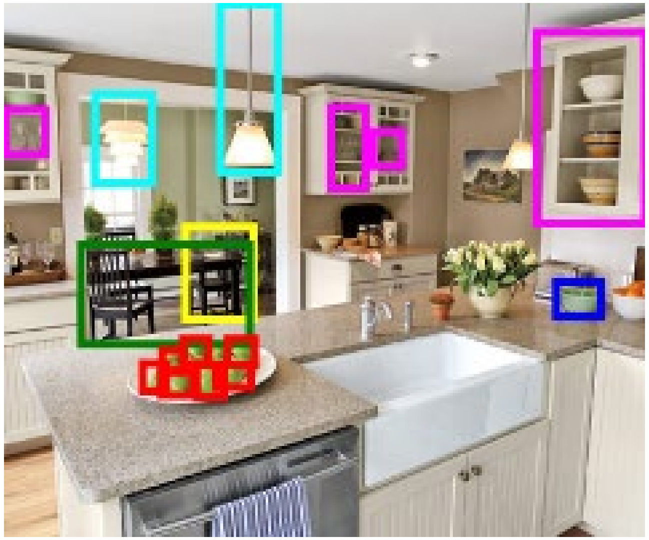
- **Object localization**
- For a given image, the algorithm produces a **class label** and a **bounding box**
- Evaluation: label that best matches the ground truth label for the image, and bounding box that overlaps with the ground truth
- An error if predicted label does not match the ground truth, or the predicted bounding box has less than 50% overlap



sea lion

Image classification is fundamental to many computer vision tasks

- **Object detection**
- Given an image, an algorithm produces a set of annotations (ci,si,bi): class label ci, bounding box bi and confidence score si
- Penalize: objects in the image not annotated by algorithm, more than 1 annotations for the same object in the image



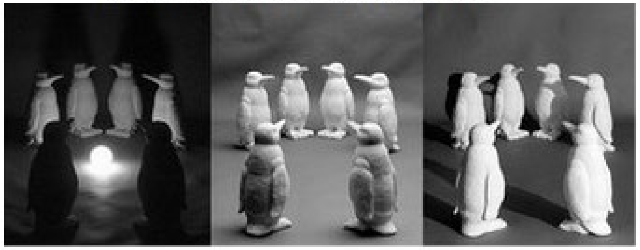
- apple
- table
- bowl
- plate rack
- lamp
- chair

200 categories in ILSVRC2017

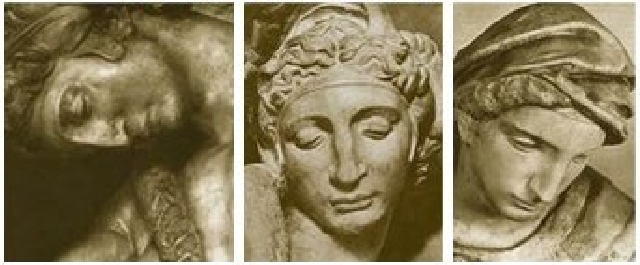
Image classification

- Challenges
  - Primitive data: Computer sees a 3d array of intensity values
  - Different variation for a certain class
    - Viewpoint variation
    - Scale variation
    - Deformation
    - Occlusion
    - Background clutter
    - Intra-class variation

Illumination conditions



Viewpoint variation



Intra-class variation

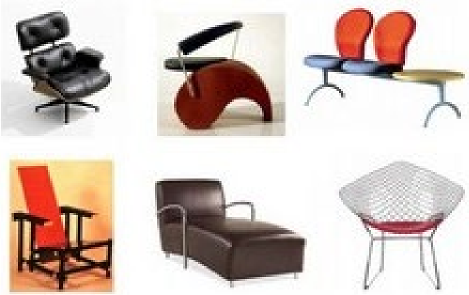


Image classification

- Challenges

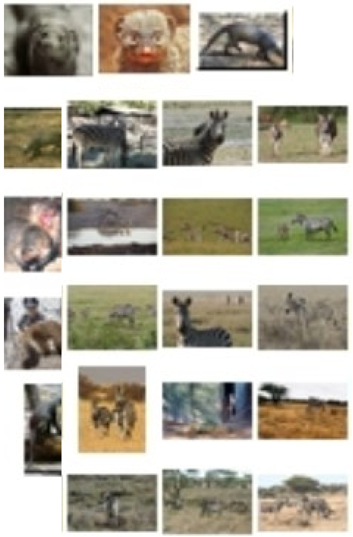
Deformation of non-rigid object

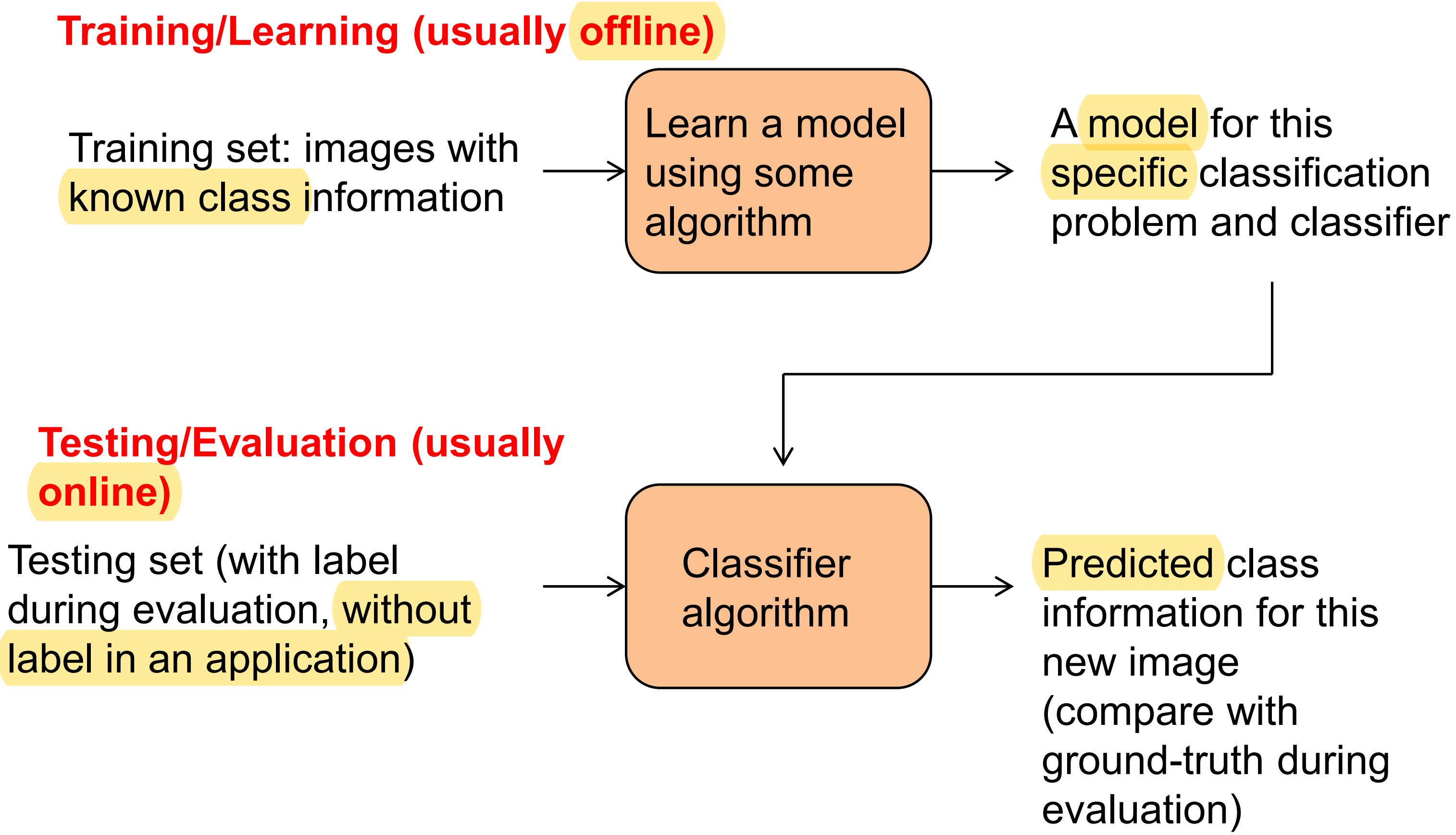
Background clutter

Data driven approach

- Provide the computer with many examples of each class: training data
- Learn the visual appearance of each class: learning algorithm
- ILSVRC: 1.2 million images of 1000 categories
  - About 1k images per category

Data driven approach





Training/Learning (usually offline)

Training set: images with known class information

Learn a model using some algorithm

A model for this specific classification problem and classifier

Testing/Evaluation (usually online)

Testing set (with label during evaluation, without label in an application)

Classifier algorithm

Predicted class information for this new image (compare with ground-truth during evaluation)

Nearest Neighbor Classifier

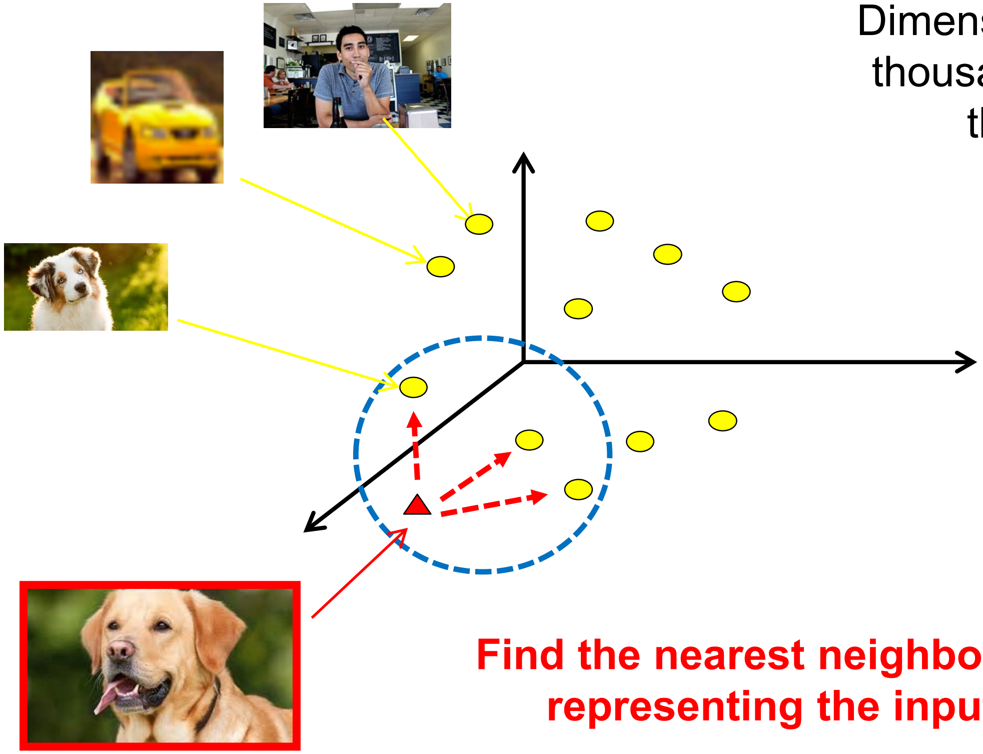
- Given a test image, compare to every one of the training images
- Use the label of the ‘closest training image’ as the predicted label

Nearest Neighbor Classifier

- Consider an image as a vector (data point) in a very high dimensional vector space
- 512x512x3 => a data point in the 786432-dim vector space
- Find the nearest neighbors of the vector representing the input test image

Nearest Neighbor Classifier

Each training image is represented by one high-dimensional vector (point)



Dimensionality can be thousands or tens of thousands

Find the nearest neighbors of the vector representing the input test image

Test image

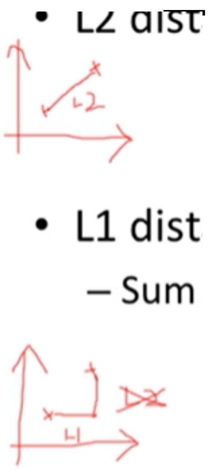
Distance

- L2 distance (Euclidean distance)

$$d_2(I_1, I_2) = \sqrt{\sum_p \left(I_1^p - I_2^p\right)^2}$$

p indicates dimension

- L1 distance (Manhattan distance)
  - Sum of abs difference



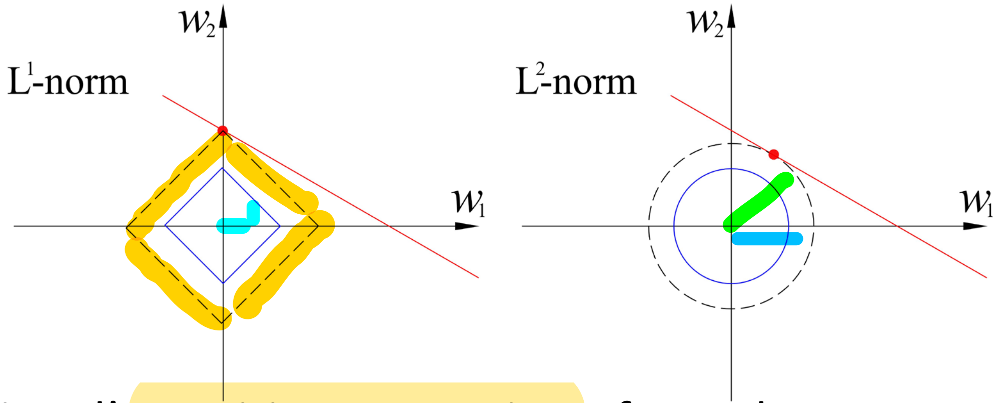
- L2 dist
- L1 dist
  - Sum

$$d_1(I_1, I_2) = \sum_p \left|I_1^p - I_2^p\right|$$

p indicates dimension

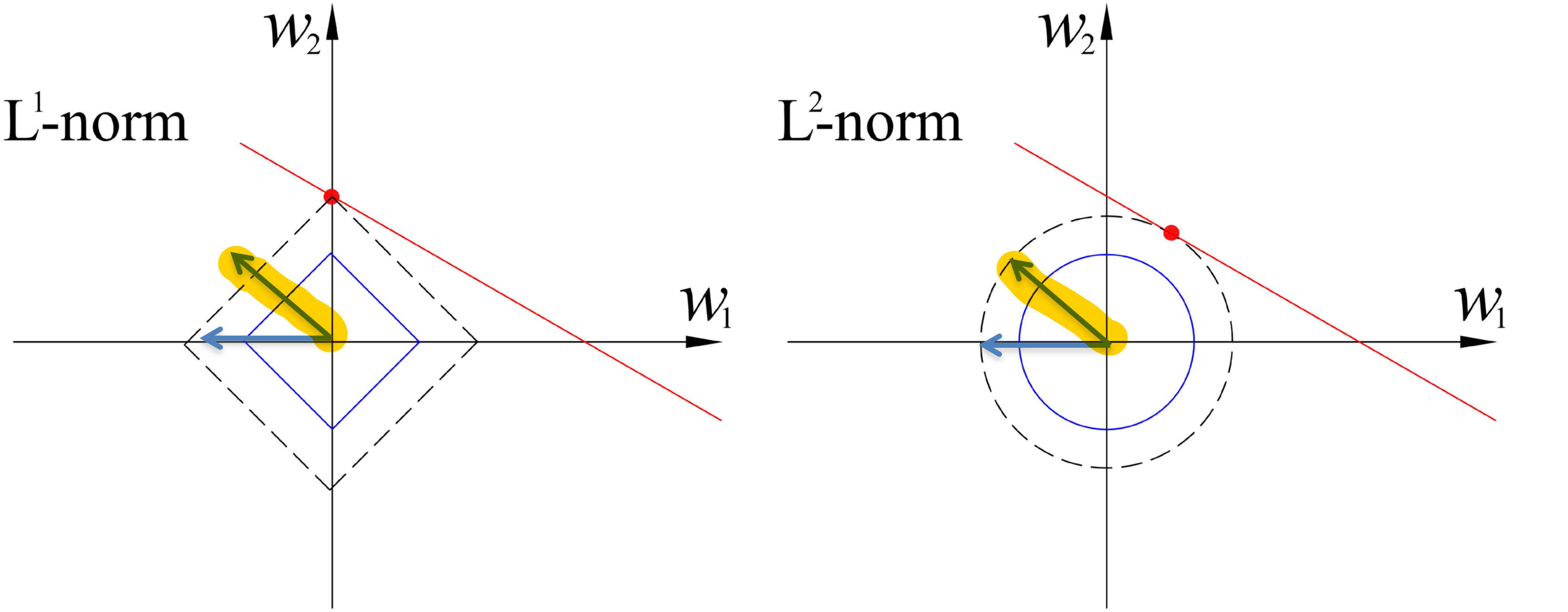
Distance

- L1/L2 circle / ball
- A circle is a set of points with a fixed distance from a point (center)



L1 is more ‘restricted’, sensitive to rotation of coord system
L2 emphasizes dimensions with large differences
L1: sparse model (use as regularization), robust to outliers (use as cost function)

Distance



L1 is more ‘restricted’, sensitive to rotation of coord system

k-Nearest Neighbor Classifier (k-NN)

- Find the k closest images (nearest neighbors)

- Use them to vote on the label of the test image

k-Nearest Neighbor Classifier (k-NN)

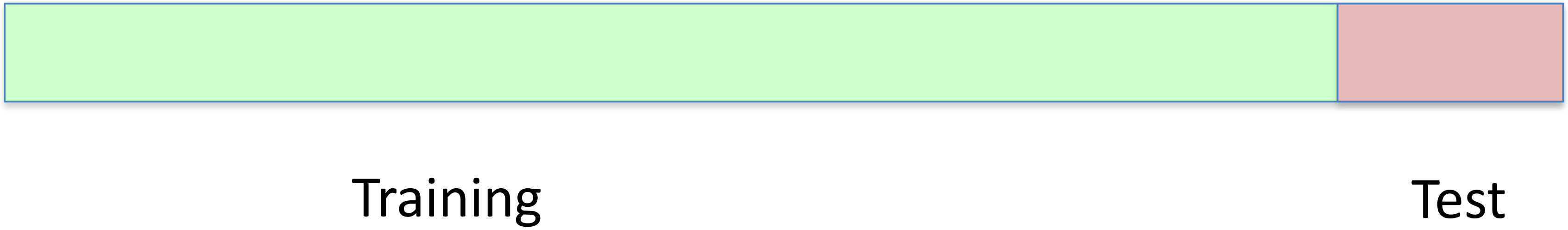
- How to determine k?

- k is a hyperparameter: related to the design of the machine learning algorithm

- Another hyperparameter: L1 norm or L2 norm

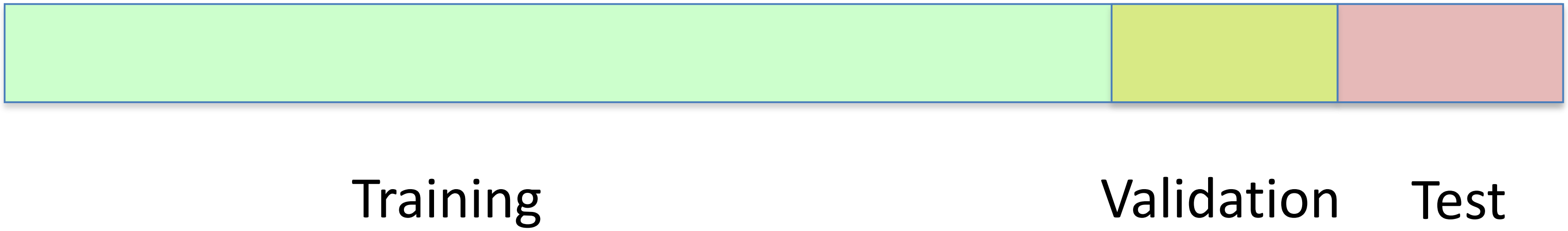
Validation set for hyperparameter tuning

- Use test set to tune the hyperparameter
- Not appropriate, as your model will overfit to the test data
- Poor generalization, significant degradation during deployment / testing for other datasets



Validation set for hyperparameter tuning

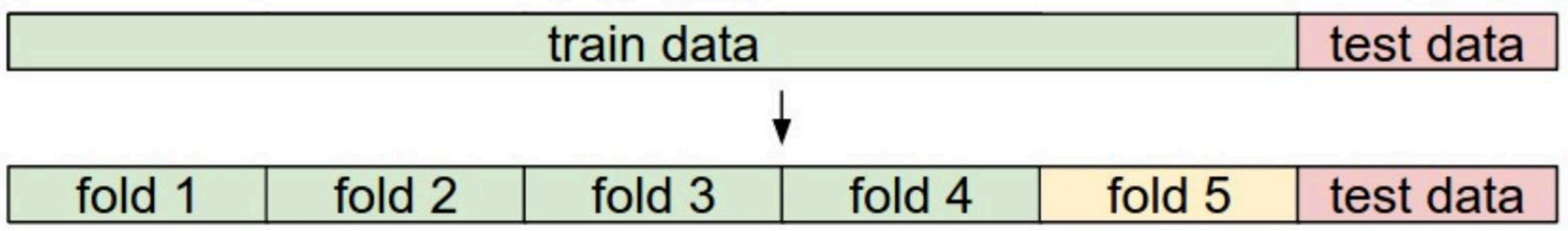
- Partition the training set into a training set and a validation set
- Use validation set to tune the hyperparameter
- Use test set to evaluate the performance



Training Validation Test

Cross validation

- If the training dataset is small, can use cross validation
- 5-fold cross validation
  - For a given k (a certain setting of hyperparameters)
  - Divide the training dataset into 5 equal folds
  - Use 4 folds for training, 1 for validation
  - Repeat using another fold as the validation set
  - Average the performance



Issues of k-NN

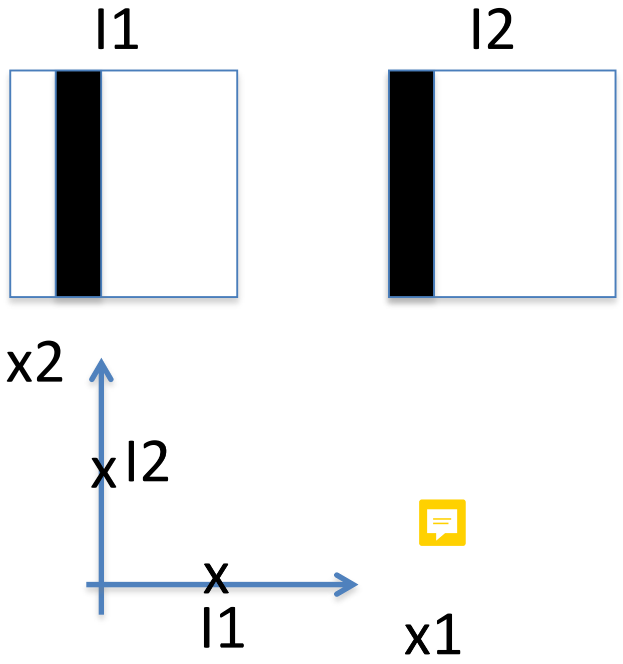
- Memory expensive: need to remember all training data
- Computationally expensive during testing
  - Need to compare all training data
  - Not practical in an application
- Approximate nearest neighbor (ANN) algorithms accelerate the search of the nearest neighbor
- Using image intensity value for distance comparison is not robust
  - Small position or intensity shift can result in large distance

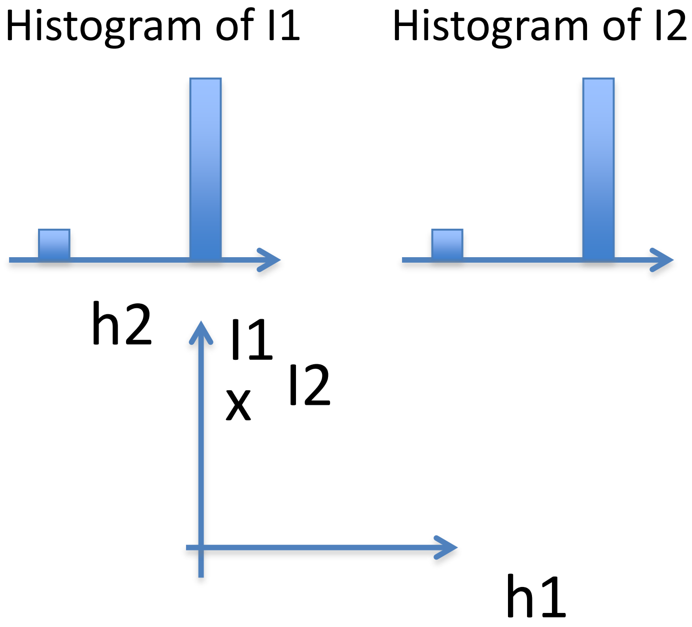
Original

Position shift

Intensity shift

Image classification: representation (feature) learning + classifier



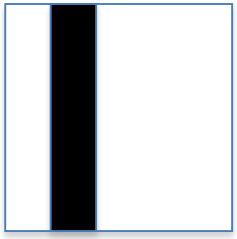


Rgb representation is sensitive to position shift (translation)

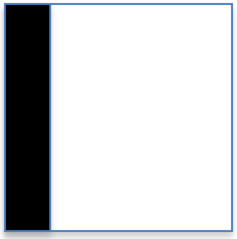
Histogram representation of an image is robust to position shift (translation)

Image classification: representation (feature) learning + classifier

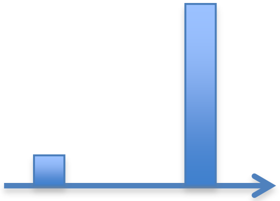
I1



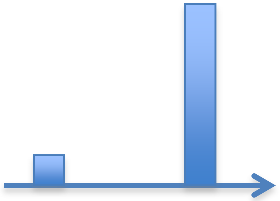
I2

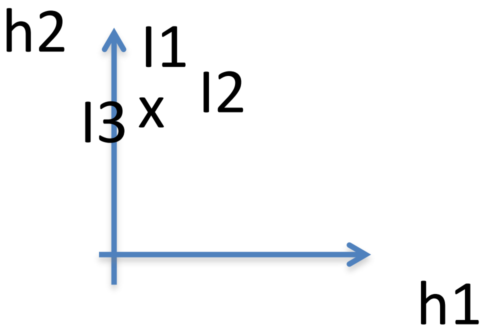


Histogram of I1

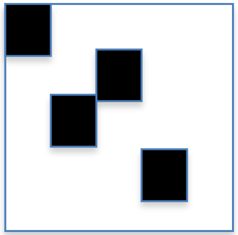


Histogram of I2

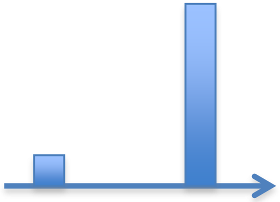




I3



Histogram of I3



Histogram representation of an image is robust to position shift (translation) but not discriminative

The quest for robust and discriminative representation