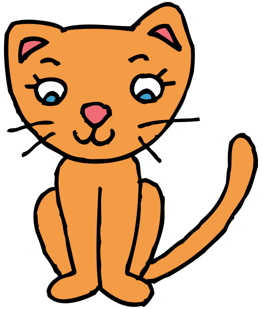
Machine Learning Basics

Machine learning

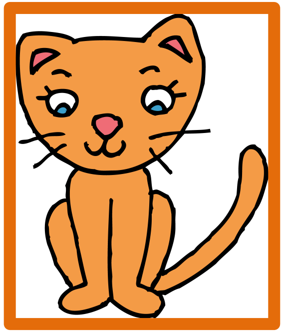
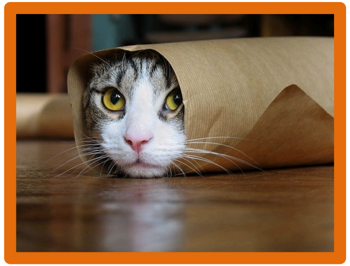


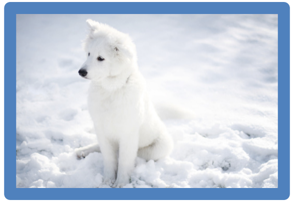


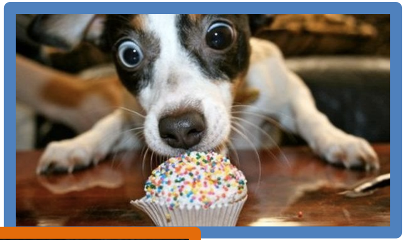

Task

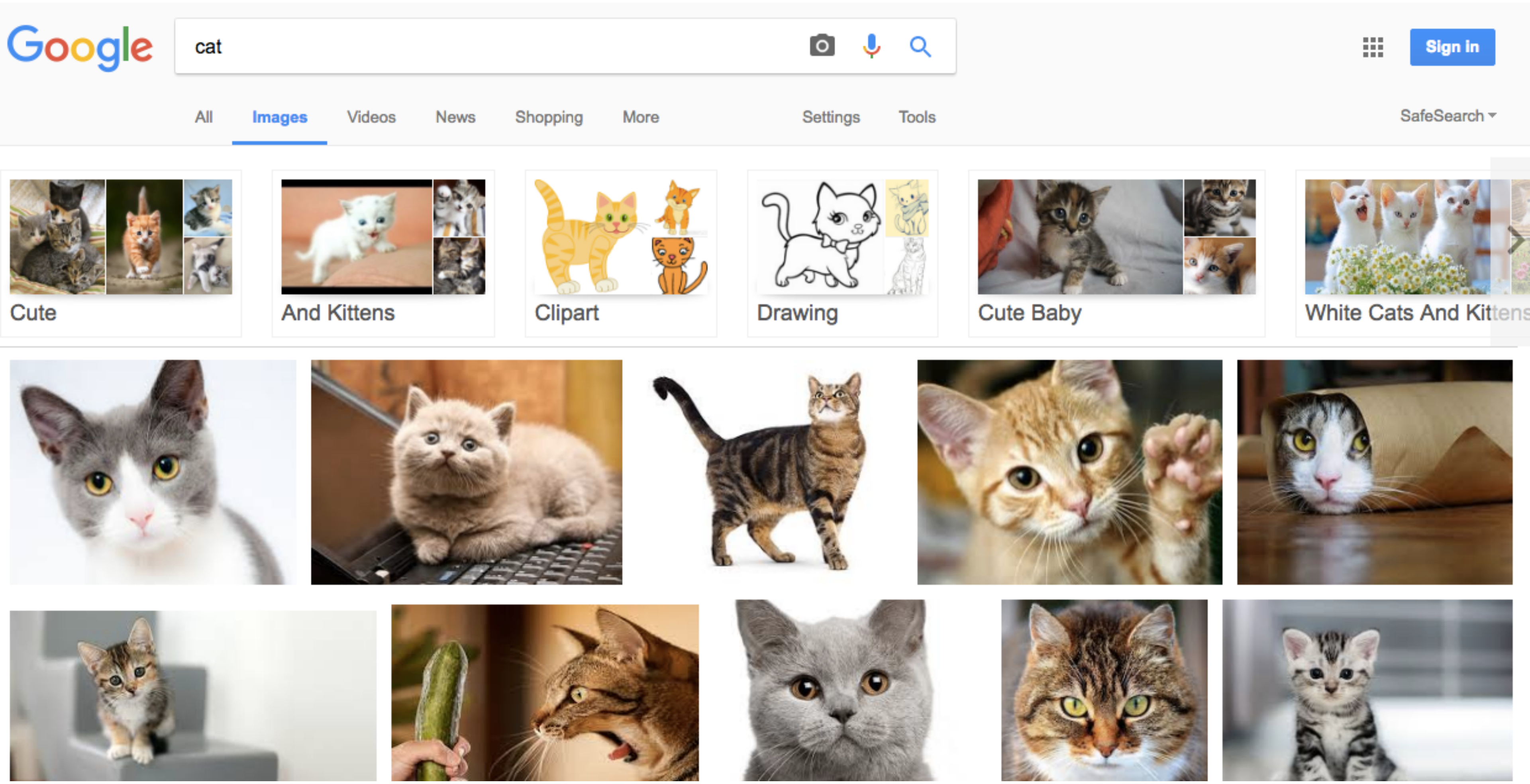



Image classification

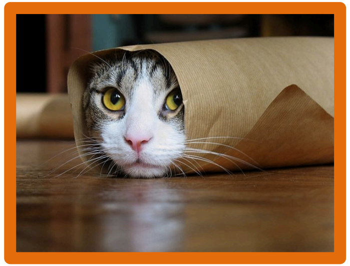



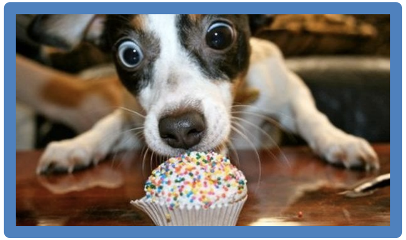





Pose Appearance Illumination

Image classification





Occlusions

Image classification

Background
clutter

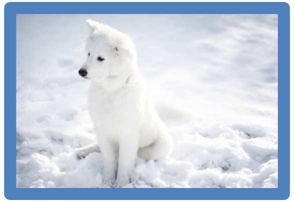
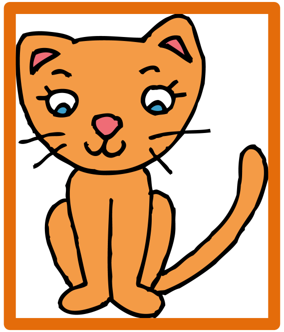


Image classification



Representation

Machine learning

- How can we learn to perform image classification?

Task

Image classification

Experience

Data

Machine learning

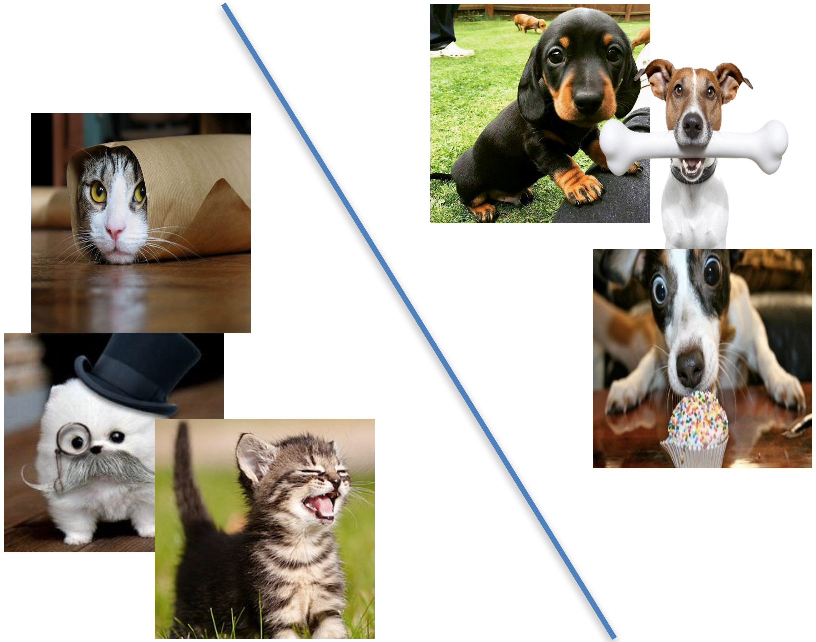
Unsupervised learning

- No label or target class
- Find out properties of the structure of the data
- Clustering (k-means, PCA)

Supervised learning

Machine learning

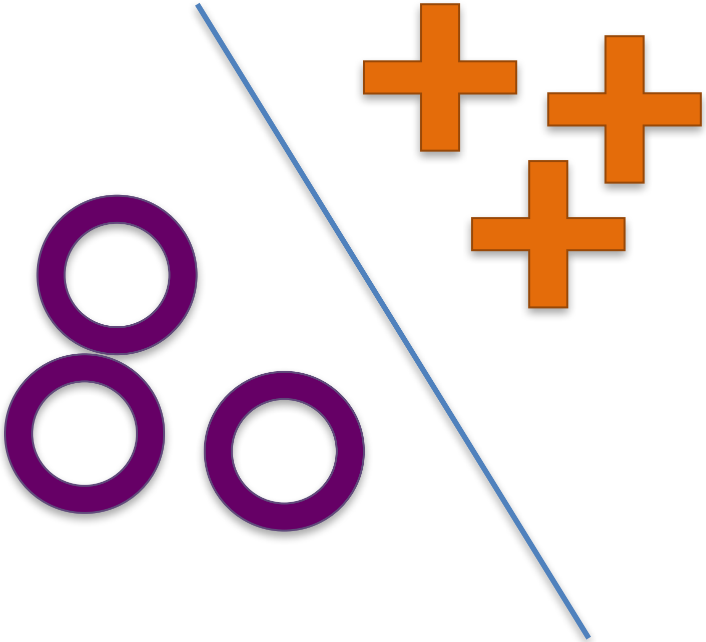
Unsupervised learning



Supervised learning

Machine learning

Unsupervised learning

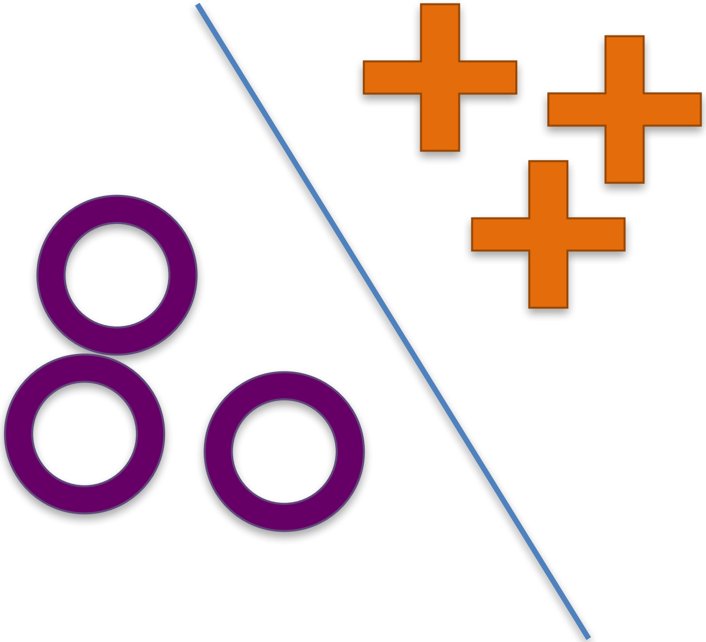


Supervised learning

- Labels or target classes

Machine learning

Unsupervised learning



Supervised learning

CAT

DOG

DOG

CAT

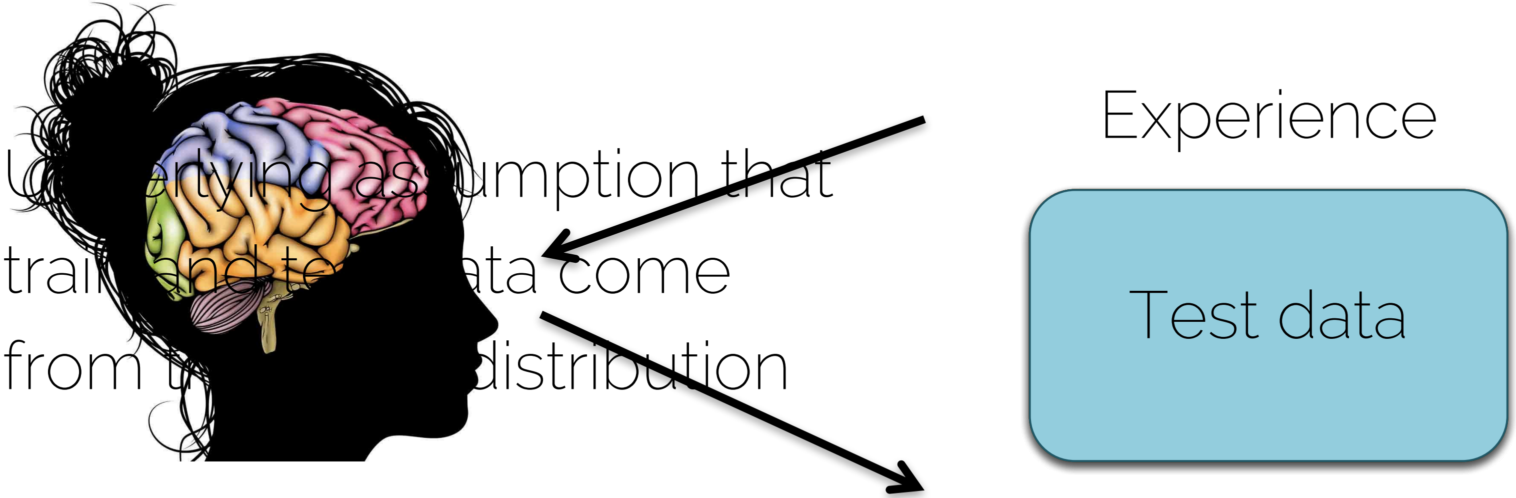
CAT

DOG

Machine learning

- How can we learn to perform image classification?

Underlying assumption that train and te[illegible] data come from t[illegible] distribution



Machine learning

Unsupervised learning

Supervised learning

Reinforcement learning

Agents → interaction → Environment

Machine learning

Unsupervised learning

Supervised learning

Reinforcement learning

Machine learning

- How can we learn to perform image classification?

Task

Image classification

Performance measure

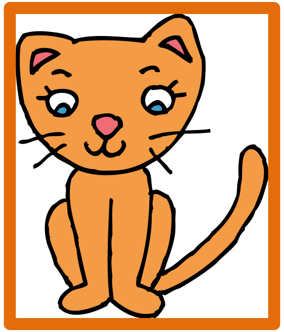
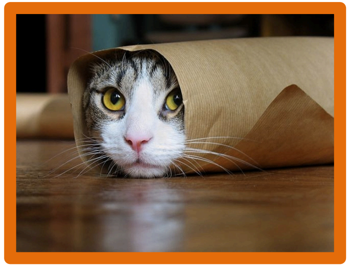
Accuracy

Experience

Data

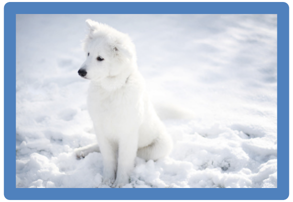
A simple classifier

Nearest Neighbor

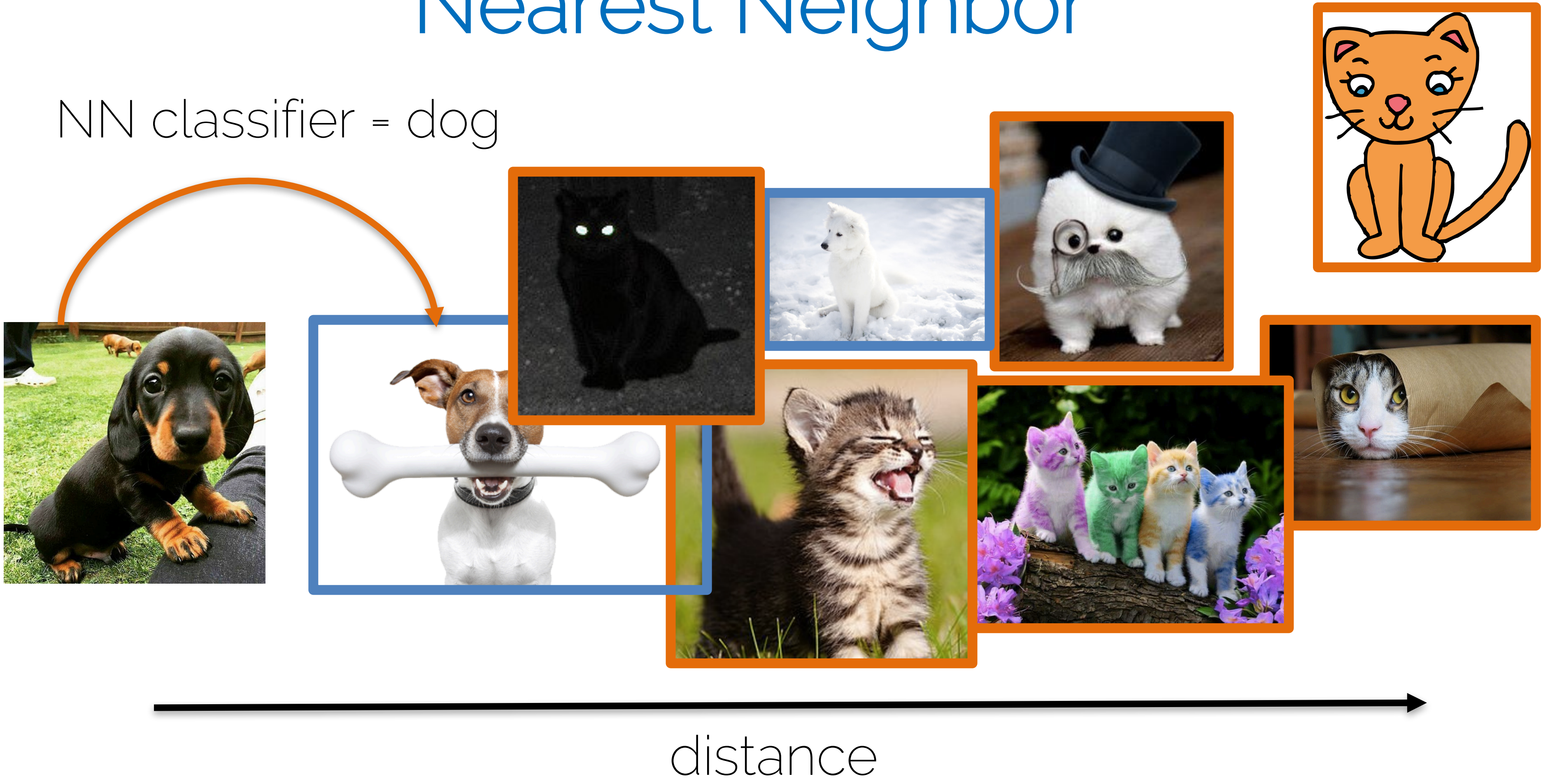







Nearest Neighbor

NN classifier = dog

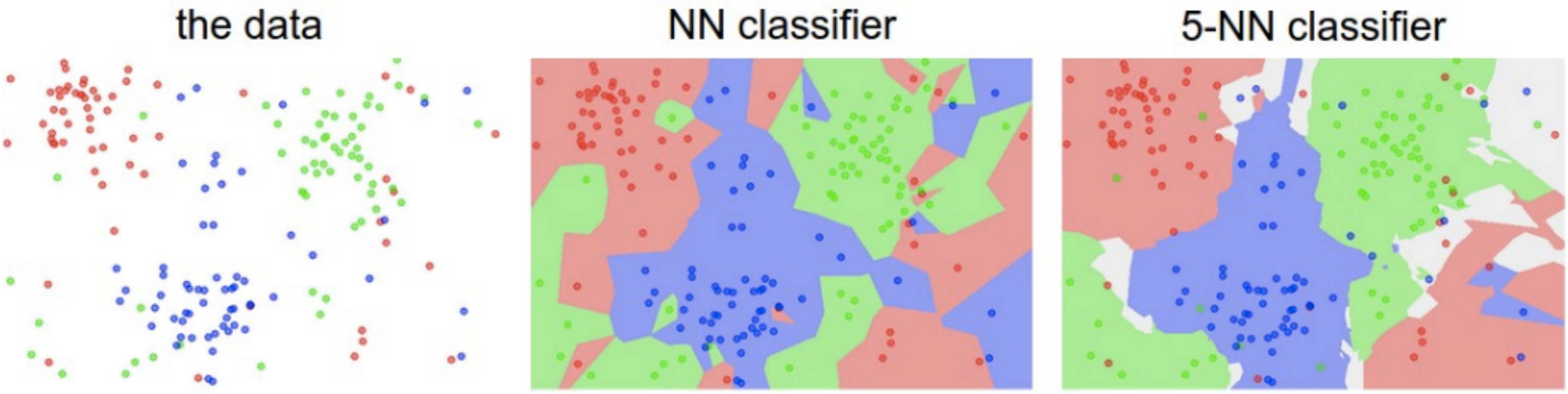


Nearest Neighbor

k-NN classifier = cat

distance

Nearest Neighbor



What is the performance on training data for NN classifier?

What classifier is more likely to perform best on test data?

Courtesy of Stanford course cs231n

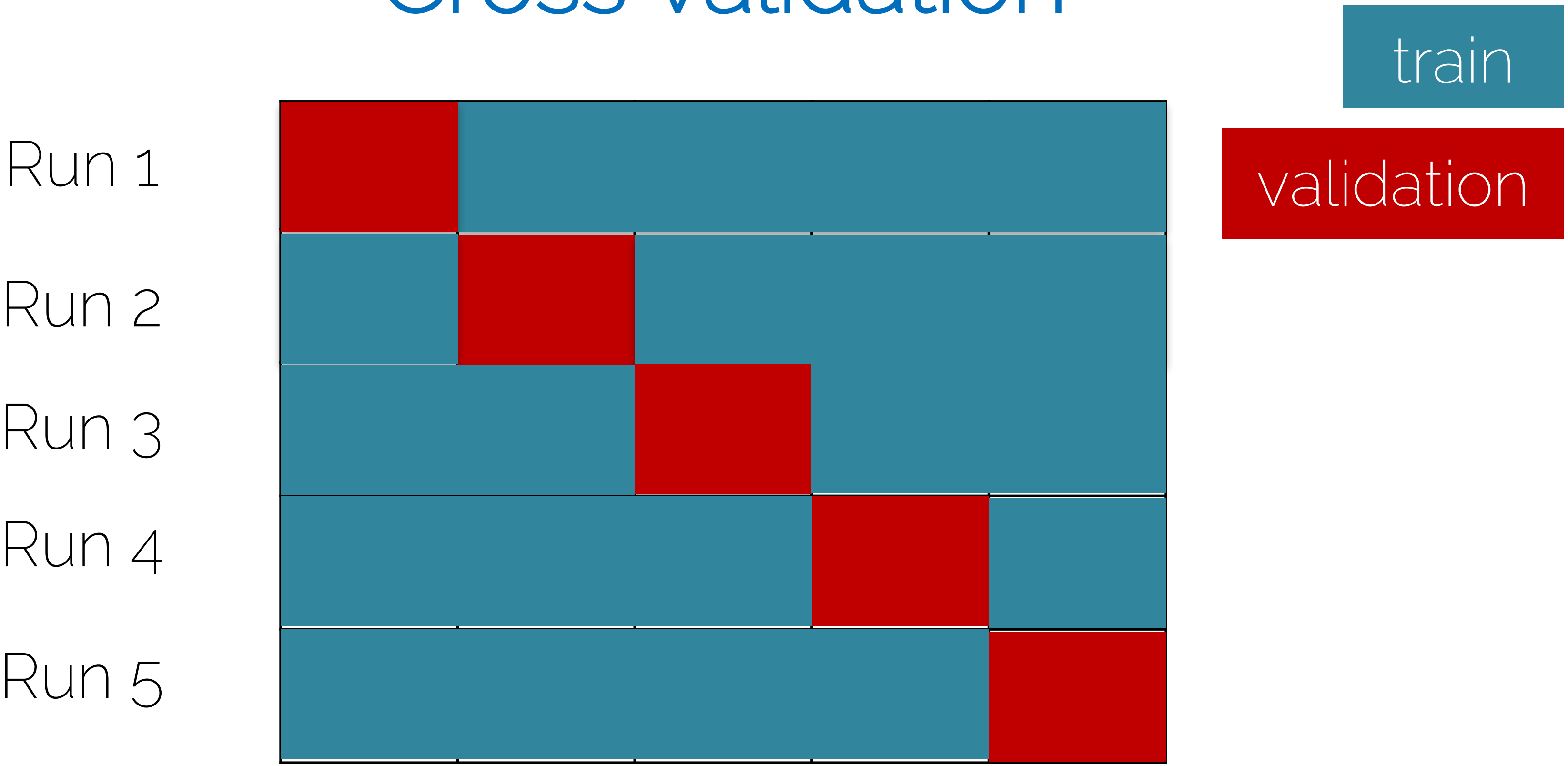
Nearest Neighbor

- Hyperparameters
  - Distance (L1, L2)
  - k (number of neighbors)

- These parameters are problem dependent.

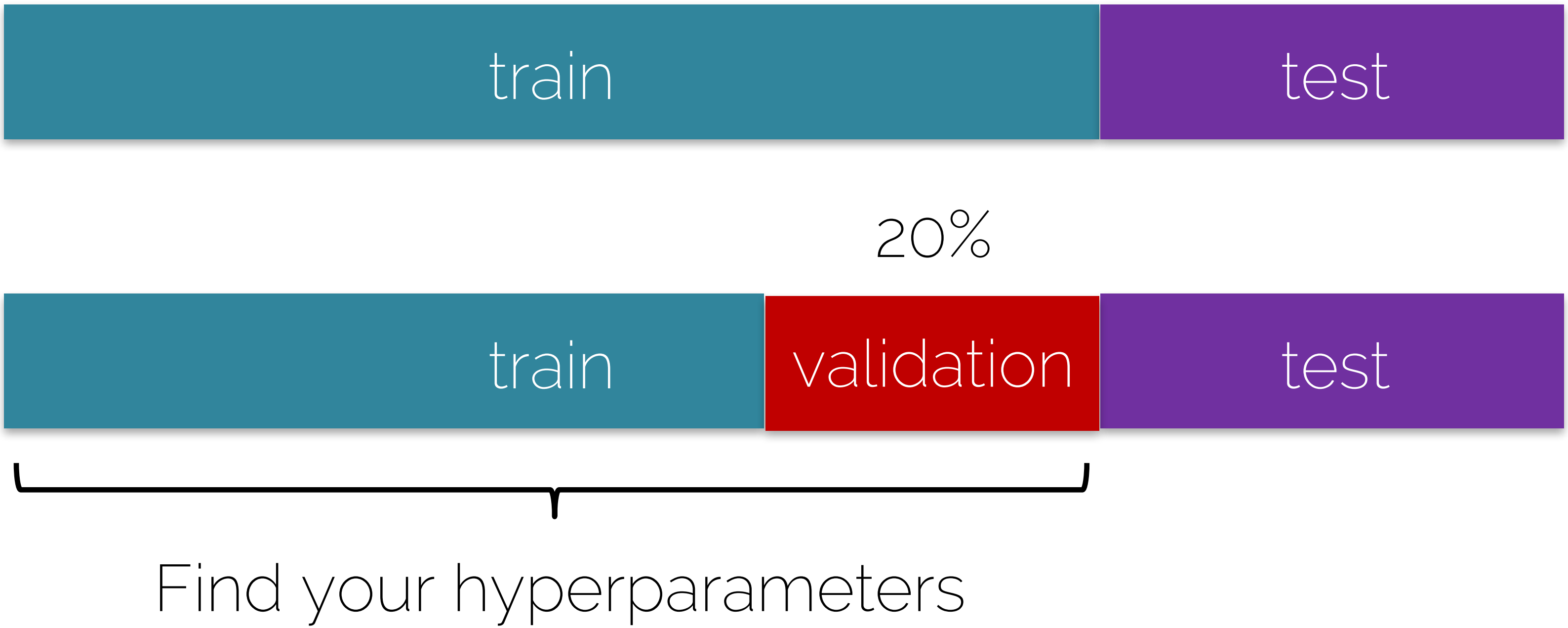
- How do we choose these hyperparameters?

Cross validation



Split the **training data** into N folds

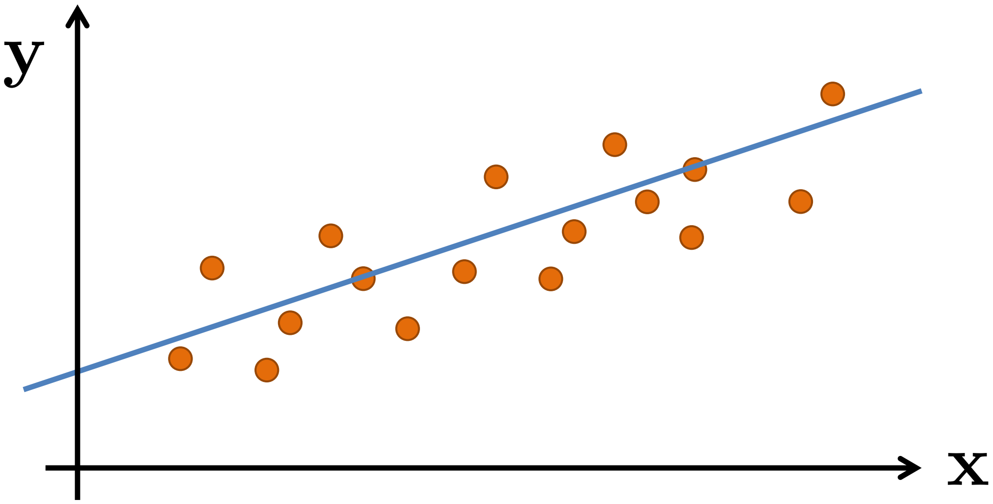
Cross validation



Linear Regression

Linear regression

- Supervised learning
- Find a linear model that explains a target $\mathbf{y}$ given the inputs $\mathbf{X}$



Linear regression

Training

$\{\mathbf{x}_{1:n}, \mathbf{y}_{1:n}\}$ $\longrightarrow$ Learner $\longrightarrow$ θ

Data points

Model parameters

Testing

$\mathbf{x}_{n+1}, \theta$ $\longrightarrow$ Predictor $\longrightarrow$ $\hat{\mathbf{y}}_{n+1}$

Estimation

Linear prediction

- A linear model is expressed in the form

d inputs

$$\hat{y}_i = \sum_{j=1}^{d} x_{ij}\theta_j$$

weights, model parameters

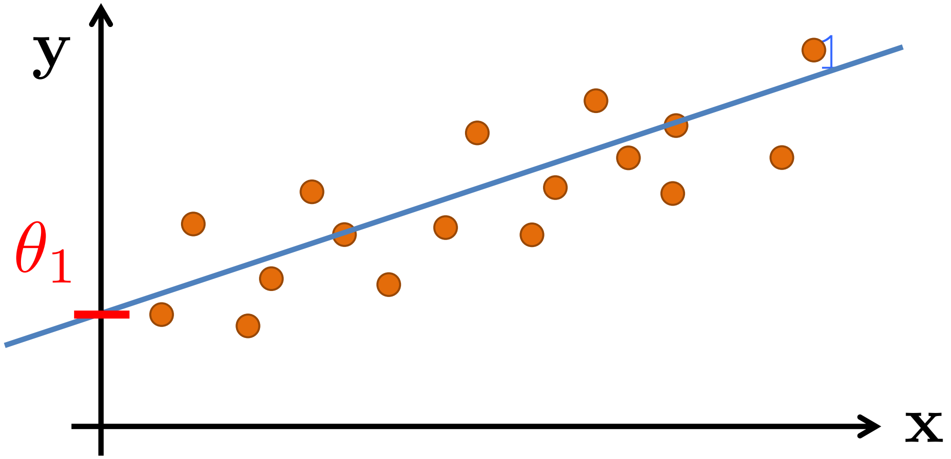
data, features

Linear prediction

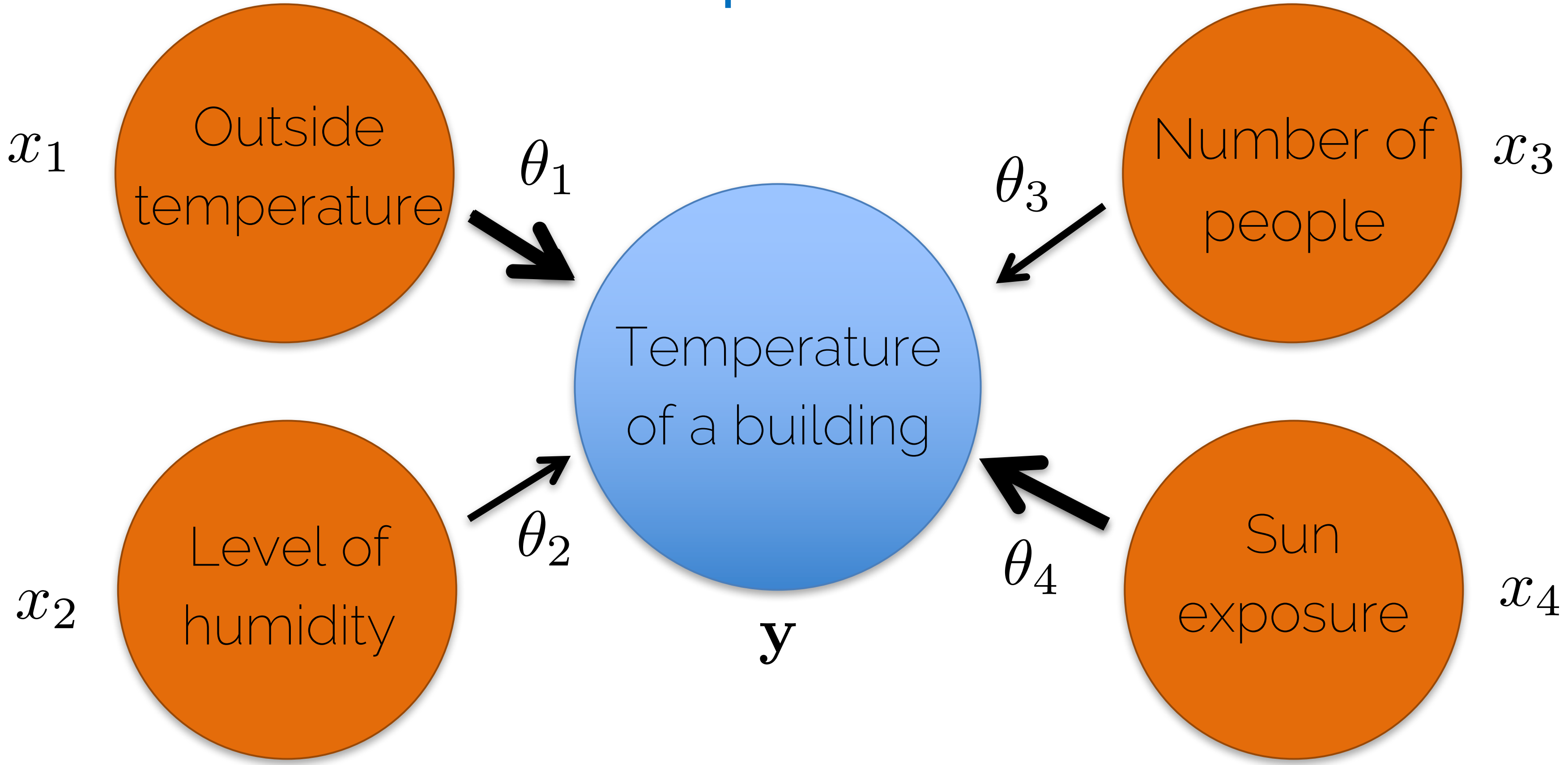
- A linear model is expressed in the form

$$\hat{y}_i = \sum_{j=1}^{d} x_{ij}\theta_j = x_{i1}\boxed{\theta_1} + x_{i2}\theta_2 + \cdots + x_{id}\theta_d$$

bias



Linear prediction



Linear prediction

$$\hat{\mathbf{y}} = \mathbf{X}\theta$$

Prediction

Input features

Model parameters

$$\begin{bmatrix} \boxed{\hat{y}_1} \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \hat{x}_{11} & \cdots & \hat{x}_{1d} \\ \hat{x}_{21} & \cdots & \hat{x}_{2d} \\ \vdots & \vdots & \vdots \\ \hat{x}_{n1} & \cdots & \hat{x}_{nd} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix}$$

Linear prediction

$$\begin{array}{c} \text{Temperature} \\ \text{of the building} \\ \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} \end{array} = \begin{array}{c} \begin{array}{cccc} \text{Outside temperature} & \text{Humidity} & \text{Number people} & \text{Sun exposure (\%)} \end{array} \\ \begin{bmatrix} 25 & 50 & 2 & 50 \\ -10 & 50 & 0 & 10 \end{bmatrix} \end{array} \begin{array}{c} \text{MODEL} \\ \begin{bmatrix} 0.64 \\ \boxed{0} \\ \boxed{1} \\ \boxed{0.14} \end{bmatrix} \end{array}$$

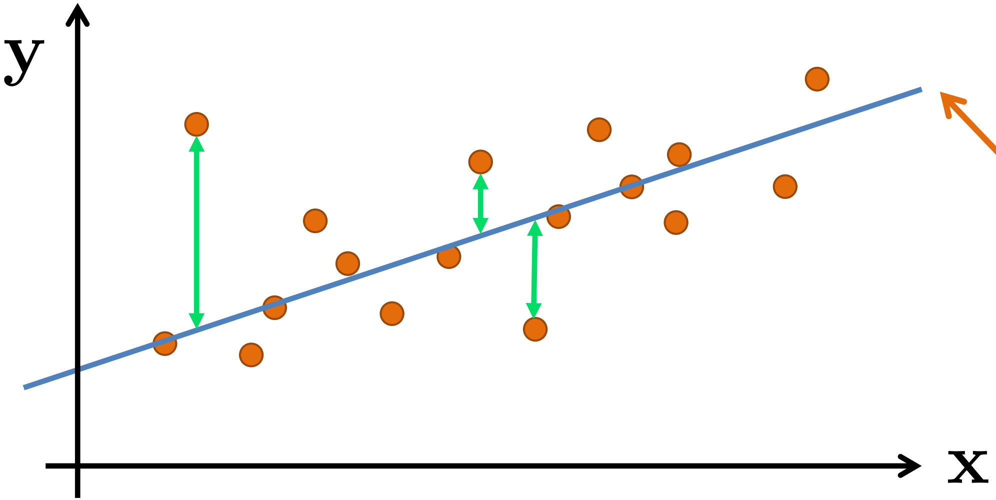
Linear prediction

How do we obtain the model?

Temperature of the building

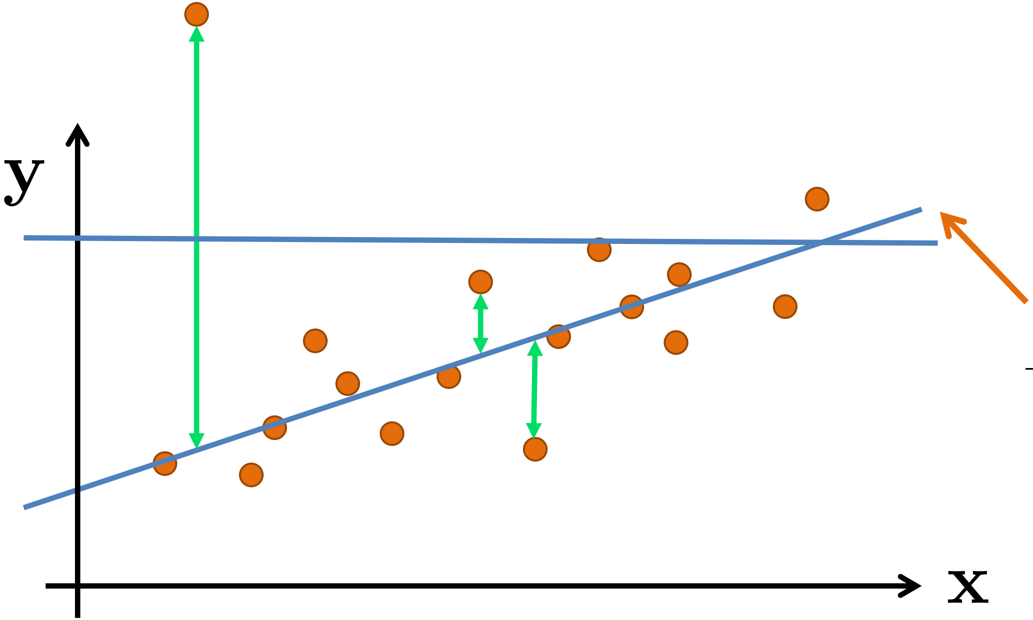
$$
\begin{array}{c} \\ \begin{bmatrix} 25 \\ -5 \end{bmatrix} \end{array}
=
\begin{array}{cccc} \text{Outside temperature} & \text{Humidity} & \text{Number people} & \text{Sun exposure (\%)} \\ \hline 25 & 50 & 2 & 50 \\ -10 & 50 & 0 & 10 \end{array}
\quad
\begin{array}{c} \text{MODEL} \\ \begin{bmatrix} 0.64 \\ 0 \\ 1 \\ 0.14 \end{bmatrix} \end{array}
$$

Linear regression



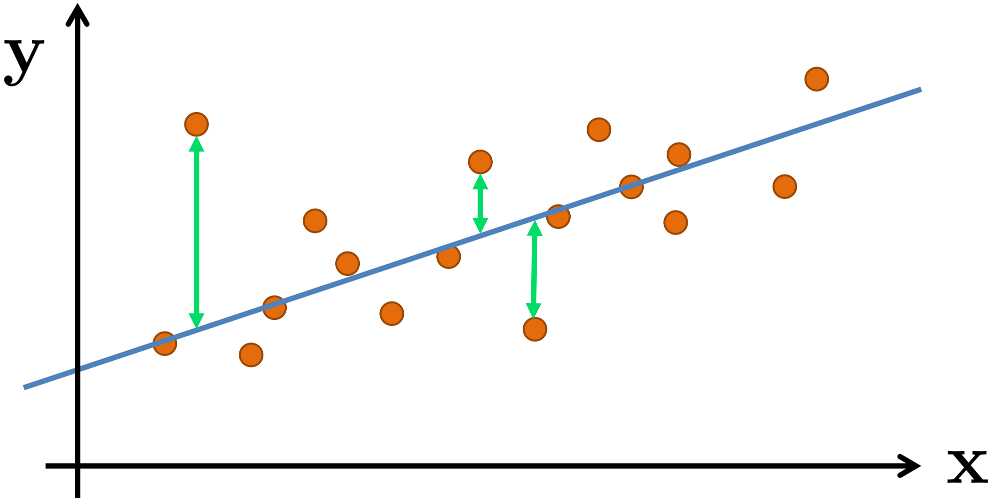
Prediction:
Temperature of
the building

Linear regression



Prediction: Temperature of the building

Linear regression



Minimizing

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Objective function

Energy

Loss

Optimization

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i\boldsymbol{\theta} - y_i)^2$$

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})$$

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

Optimization

$$J(\boldsymbol{\theta}) = (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})$$

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{y} + \cancel{\mathbf{y}^T\mathbf{y}})$$

$$\frac{\partial \boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}}{\partial \boldsymbol{\theta}} = 2\mathbf{A}^T\boldsymbol{\theta}$$

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 2\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - 2\mathbf{X}^T\mathbf{y} = 0$$

Optimization

$$\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 2\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - 2\mathbf{X}^T\mathbf{y} = 0$$

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Inputs: Outside temperature, number of people…

Output: Temperature of the building

Is this the best estimate?

- Least squares estimate

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Maximum Likelihood

Maximum Likelihood Estimate

$p_{data}(\mathbf{x})$ True underlying distribution

$p_{model}(\mathbf{x}; \boldsymbol{\theta})$ Parametric family of distributions

Controlled by a parameter

Maximum Likelihood Estimate

- A method of estimating the parameters of a statistical model given observations,

$$p_{model}(\mathbb{X};\boldsymbol{\theta})$$

Observations from $p_{data}(\mathbf{x})$

Maximum Likelihood Estimate

- A method of estimating the parameters of a statistical model given observations, by finding the parameter values that **maximize the likelihood** of making the observations given the parameters.

$$\boldsymbol{\theta}_{ML} = \arg\max_{\theta} p_{model}(\mathbb{X}; \boldsymbol{\theta})$$

$$\boldsymbol{\theta}_{ML} = \arg\max_{\theta} \prod_{i=1}^{m} p_{model}(\mathbf{x}_i; \boldsymbol{\theta})$$

Maximum Likelihood Estimate

$$\boldsymbol{\theta}_{ML} = \arg\max_{\theta} \prod_{i=1}^{m} p_{model}(\mathbf{x}_i; \boldsymbol{\theta})$$

$$\boldsymbol{\theta}_{ML} = \arg\max_{\theta} \sum_{i=1}^{m} \log p_{model}(\mathbf{x}_i; \boldsymbol{\theta})$$

Spoiler: Related to softmax loss

Back to linear regression

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$$

Probability distribution

Input data

Back to linear regression

$$\boldsymbol{\theta}_{ML} = \arg\max_{\theta} p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$$

i.i.d. =independent and
identically distributed

$$\boldsymbol{\theta}_{ML} = \arg\max_{\theta} \sum_{i=1}^{m} \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$$

Back to linear regression

$$\sum_{i=1}^{m} p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$$

Gaussian or Normal distribution

Assuming $\quad y_i = \mathcal{N}(\mathbf{x}_i\boldsymbol{\theta}, \sigma^2) = \mathbf{x}_i\boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$

mean

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad x \sim \mathcal{N}(\mu, \sigma^2)$$

Back to linear regression

$$\sum_{i=1}^{m} p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$$

Assuming $\quad y_i = \mathcal{N}(\mathbf{x}_i\boldsymbol{\theta}, \sigma^2) = \mathbf{x}_i\boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad x \sim \mathcal{N}(\mu, \sigma^2)$$

Back to linear regression

$$\sum_{i=1}^{m} p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i\boldsymbol{\theta})^2}$$

Assuming $y_i = \mathcal{N}(\mathbf{x}_i\boldsymbol{\theta}, \sigma^2) = \mathbf{x}_i\boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad x \sim \mathcal{N}(\mu, \sigma^2)$$

Back to linear regression

$$\sum_{i=1}^{m} p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i\boldsymbol{\theta})^2}$$

Assuming $\quad y_i = \mathcal{N}(\mathbf{x}_i\boldsymbol{\theta}, \sigma^2) = \mathbf{x}_i\boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad x \sim \mathcal{N}(\mu, \sigma^2)$$

Back to linear regression

$$\sum_{i=1}^{m} p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i\boldsymbol{\theta})^2}$$

$$\boldsymbol{\theta}_{ML} = \arg\max_{\theta} \sum_{i=1}^{m} \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$$

Back to linear regression

$$\log((2\pi\sigma^2)^{-1/2}e^{-\frac{1}{2\sigma^2}(y_i-\mathbf{x}_i\boldsymbol{\theta})^2})$$

$$\Big\downarrow \text{Matrix notation}$$

$$\log((2\pi\sigma^2)^{-1/2}e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})})$$

$$\Big\downarrow$$

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$

Back to linear regression

$$\boldsymbol{\theta}_{ML} = \arg\max_{\theta} p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})$$

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$

$$\Big\downarrow \frac{\partial}{\partial \boldsymbol{\theta}}$$

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

How can we find the estimate of theta?

Back to linear regression

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$

Can you derive the estimate of sigma?

Regularization and MAP

Regularization

$x = [1, 2, 1]$ → Input = 3 features

$\theta_1 = [1.5, 0, 0]$ → Ignores 2 features

$\theta_2 = [0.25, 0.5, 0.25]$ → Takes information from all features

Regularization

Loss $$J(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \boxed{\lambda R(\boldsymbol{\theta})}$$

L2 regularization $$\boldsymbol{\theta}^T\boldsymbol{\theta}$$

$$\boldsymbol{\theta}_1^T\boldsymbol{\theta}_1 = 1.5 * 1.5 = 2.25$$

$$\boldsymbol{\theta}_2^T\boldsymbol{\theta}_2 = 0.25^2 + 0.5^2 + 0.25^2 = \boxed{0.375}$$

$x = [1, 2, 1]$ $\theta_1 = [1.5, 0, 0]$ $\theta_2 = [0.25, 0.5, 0.25]$

Regularization

Loss $$J(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + \lambda R(\boldsymbol{\theta})$$

L2 regularization

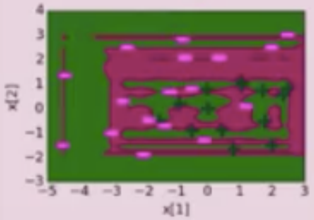
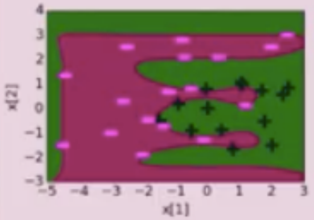
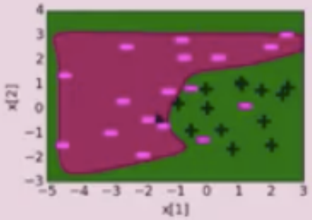
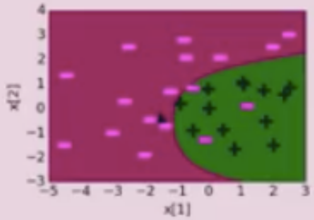
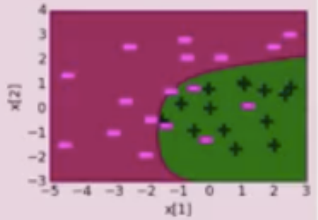
L1 regularization

Max norm regularization

Dropout

Can you find the relationship between this loss and the Maximum a Posteriori (MAP) estimate?

Regularization

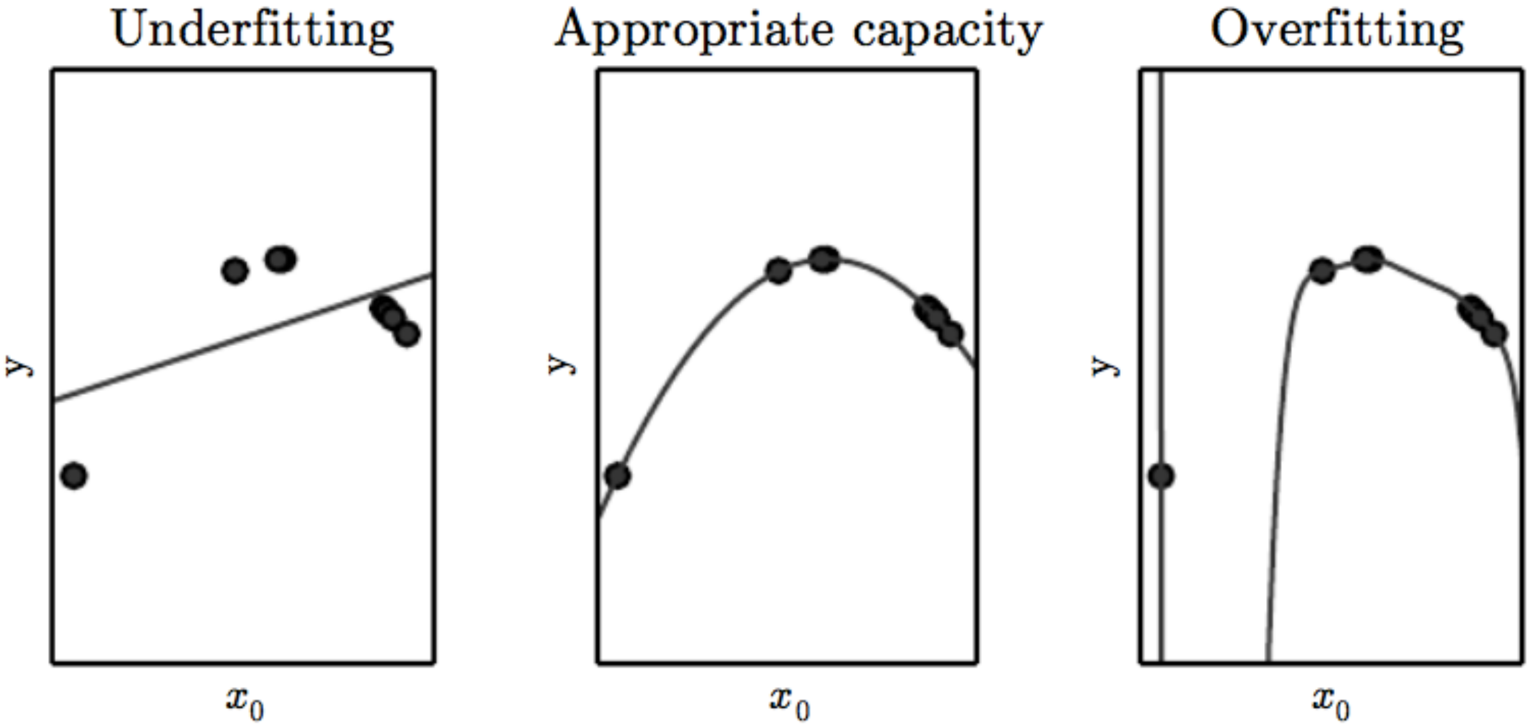
Regularization	λ = 0	λ = 0.00001	λ = 0.001	λ = 1	λ = 10
Decision boundary					

What is the goal of regularization?

What happens to the training error?

Credits: University of Washington

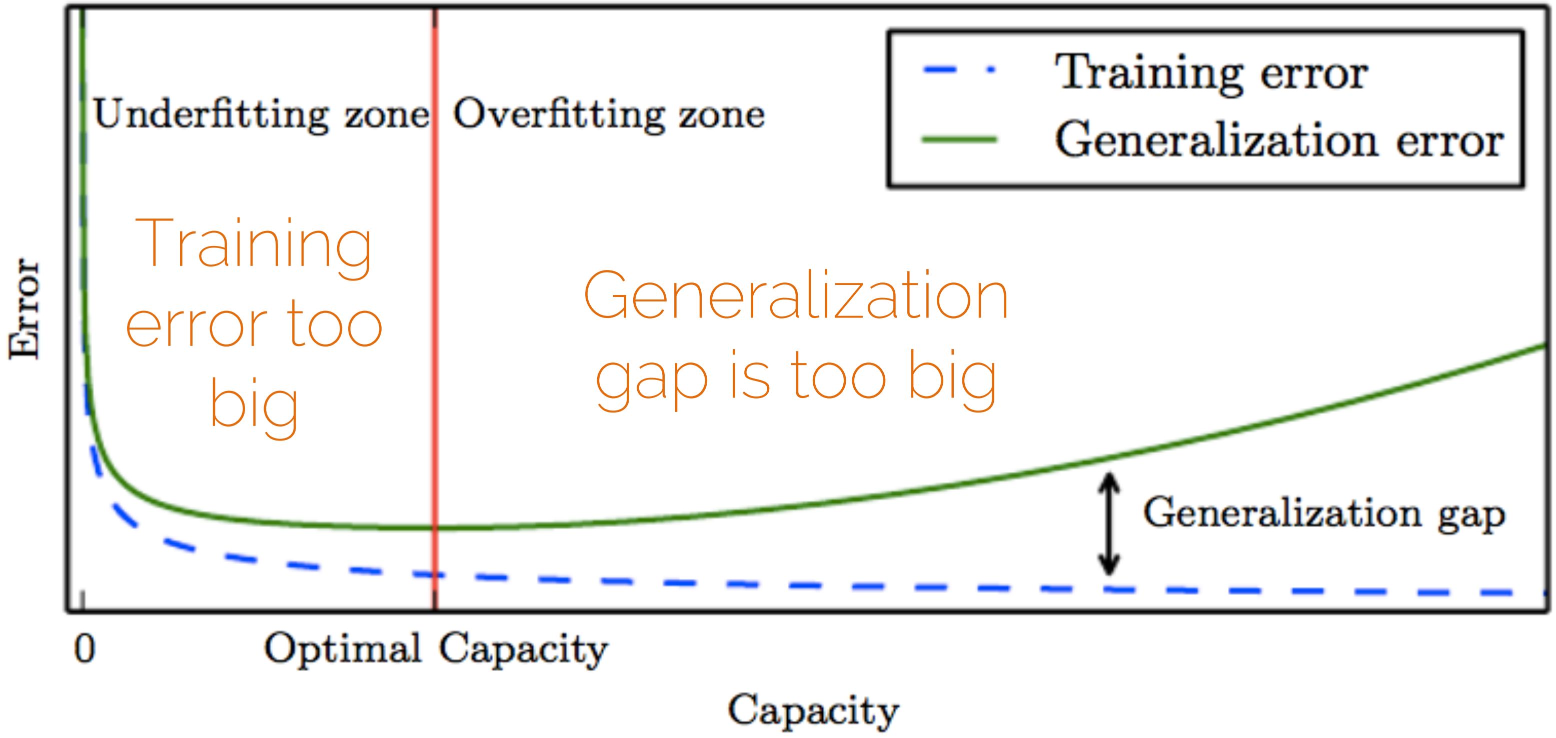
Overfitting and underfitting



What is lambda for each of the cases?

Credits: Deep Learning. Goodfellow et al.

Overfitting and underfitting



Credits: Deep Learning. Goodfellow et al.

Visualization

http://vision.stanford.edu/teaching/cs231n-demos/linear-classify/

Next lectures

- Next Tuesday: Introduction to Neural Networks
- First exercise on Thursday 2nd of November here!