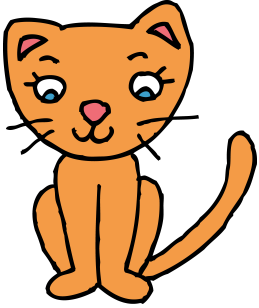


Machine Learning Basics

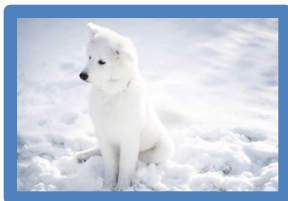
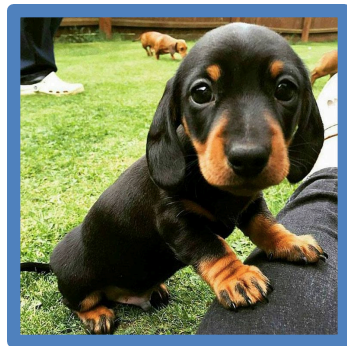
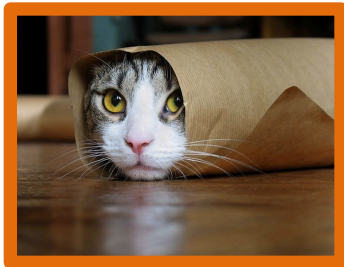
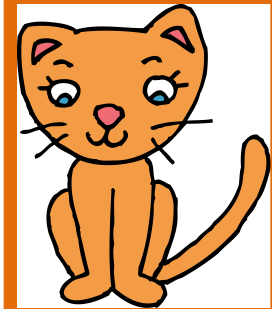
Machine learning



Task

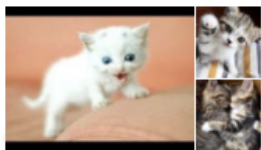


Image classification





Cute



And Kittens



Clipart



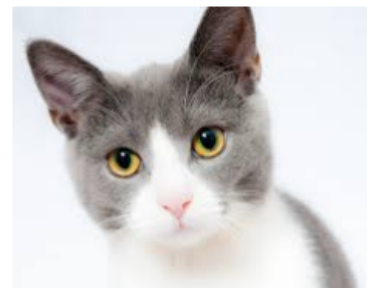
Drawing



Cute Baby



White Cats And Kittens



Pose



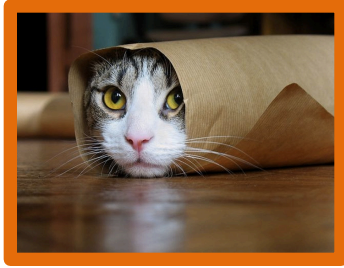
Appearance



Illumination



Image classification

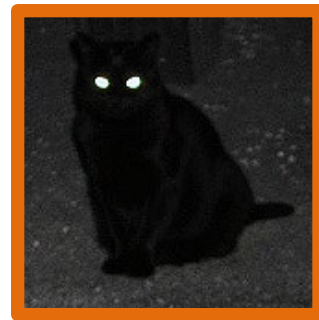
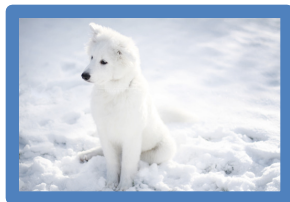


Occlusions



Image classification

Background
clutter



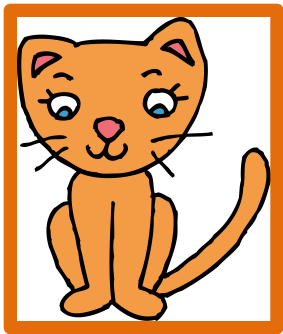


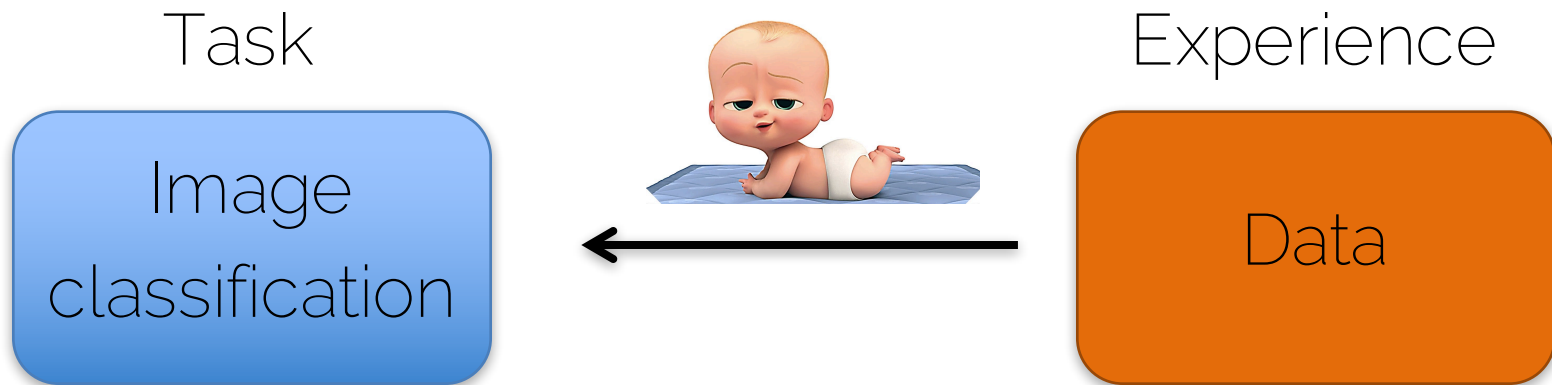
Image classification

Representation



Machine learning

- How can we learn to perform image classification?



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

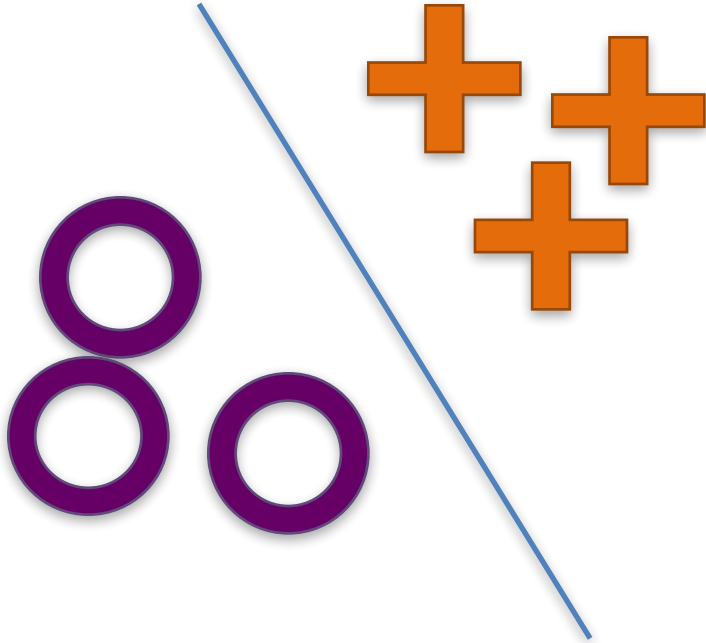


Supervised learning



Machine learning

Unsupervised learning

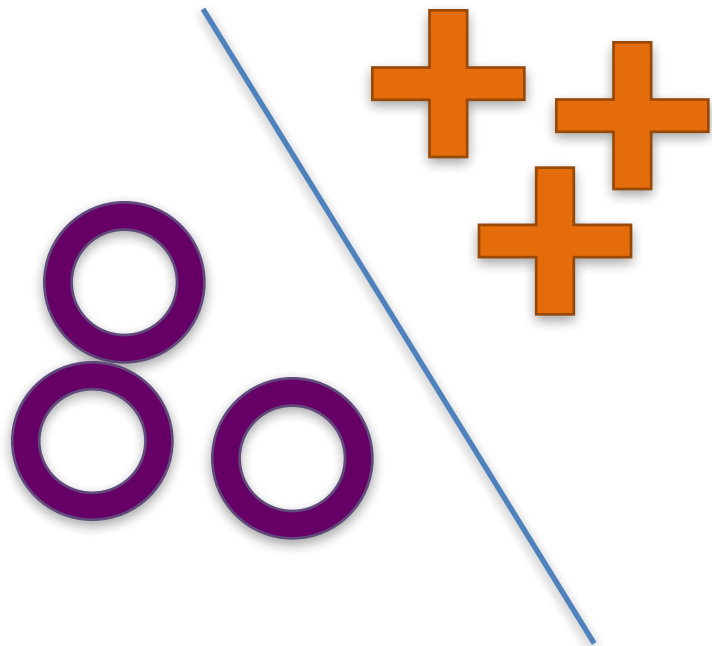


Supervised learning

- Labels or target classes

Machine learning

Unsupervised learning

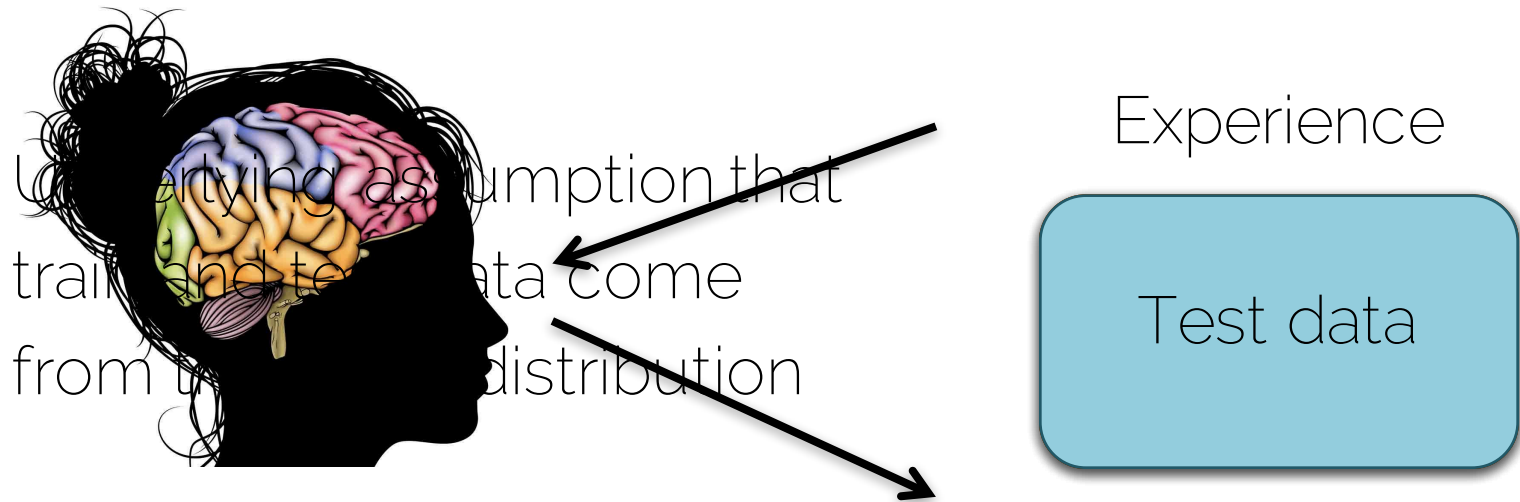


Supervised learning



Machine learning

- How can we learn to perform image classification?



Machine learning

Unsupervised learning



Supervised learning



Reinforcement learning

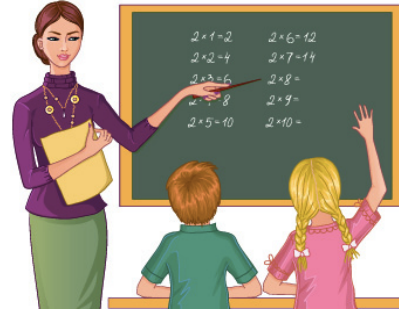


Machine learning

Unsupervised learning



Supervised learning



Reinforcement learning



reward



Machine learning

- How can we learn to perform image classification?

Task

Image
classification

Performance
measure

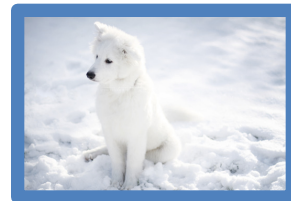
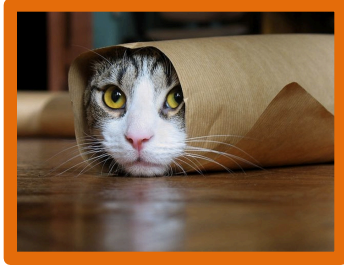
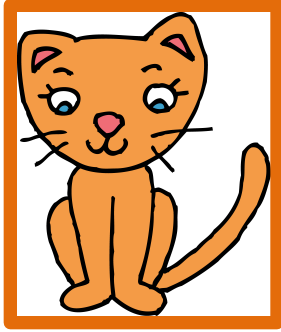
Accuracy

Experience

Data

A simple classifier

Nearest Neighbor



Nearest Neighbor

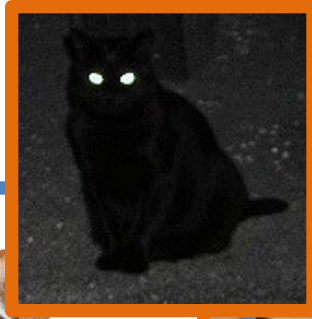
NN classifier = dog



distance

Nearest Neighbor

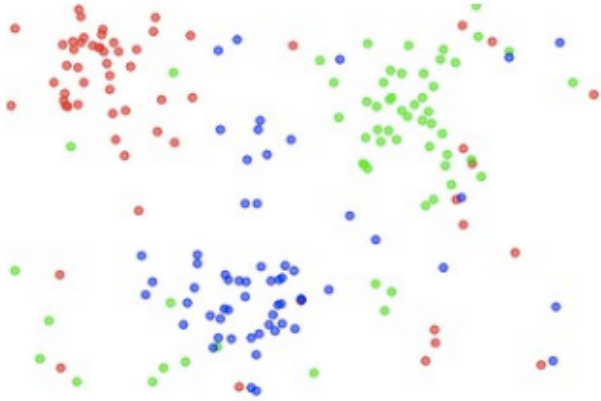
k-NN classifier = cat



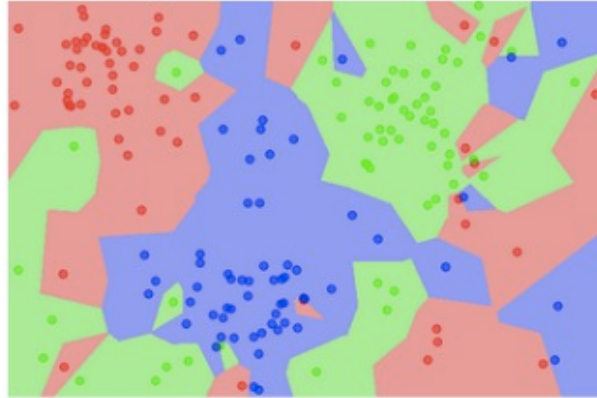
distance

Nearest Neighbor

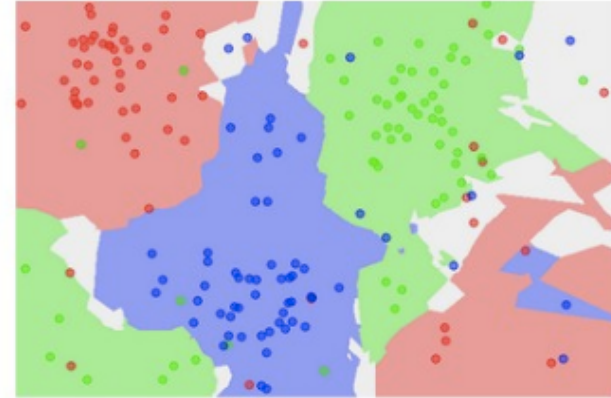
the data



NN classifier



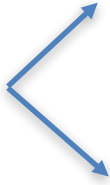
5-NN classifier



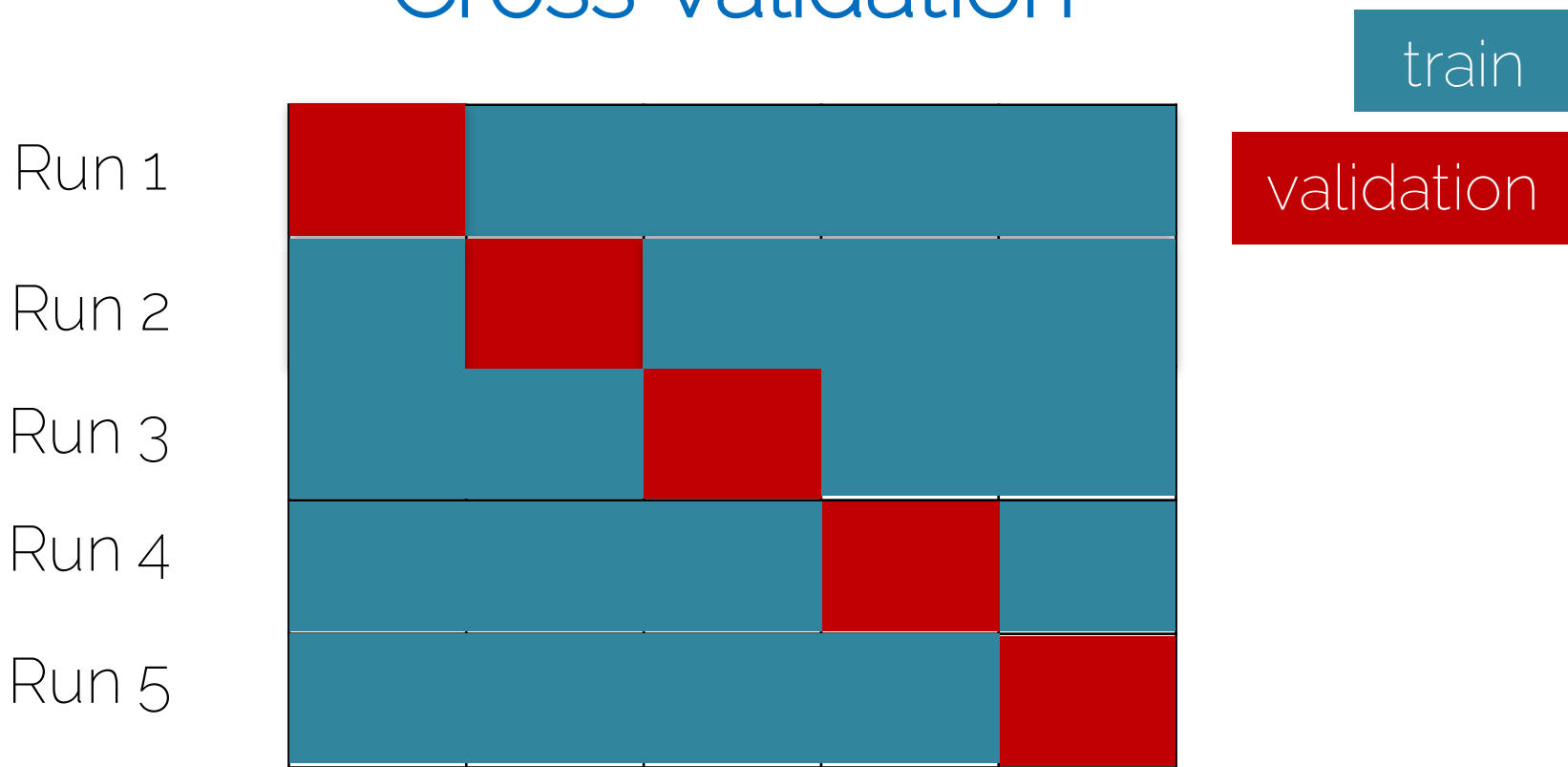
What is the performance on training data for NN classifier?

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

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

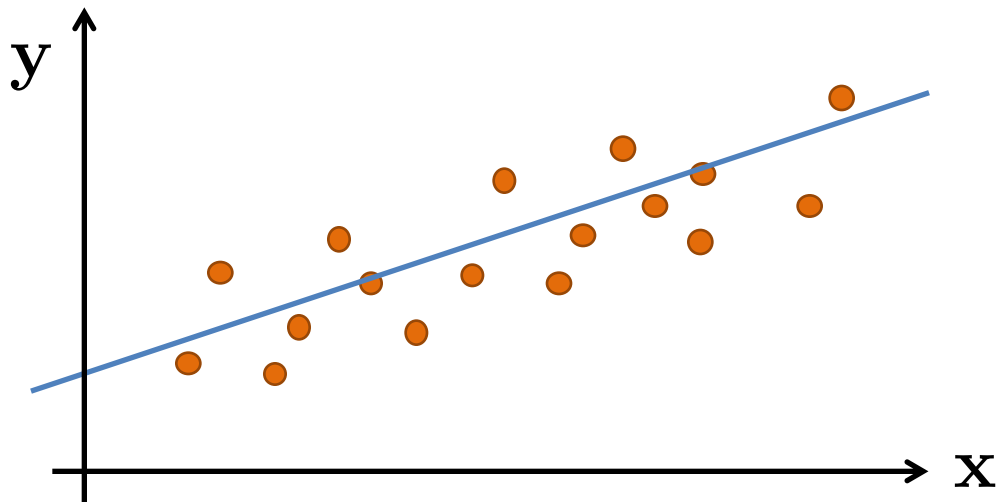


Find your hyperparameters

Linear Regression

Linear regression

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



Linear regression

Training



Testing



Linear prediction

- A linear model is expressed in the form

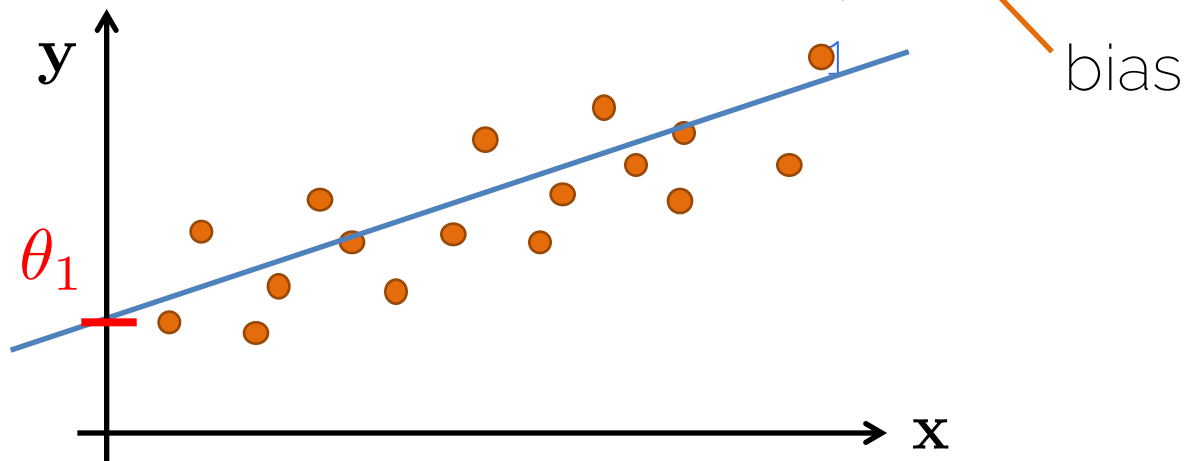
$$\hat{y}_i = \sum_{j=1}^d x_{ij} \theta_j$$

The diagram illustrates the components of the linear model equation $\hat{y}_i = \sum_{j=1}^d x_{ij} \theta_j$. The summation index d is annotated with a purple arrow pointing to the text "d input features". The term x_{ij} is circled in orange, with an orange arrow pointing to the text "data, features". The term θ_j is circled in blue, with a blue arrow pointing to the text "weights, model parameters".

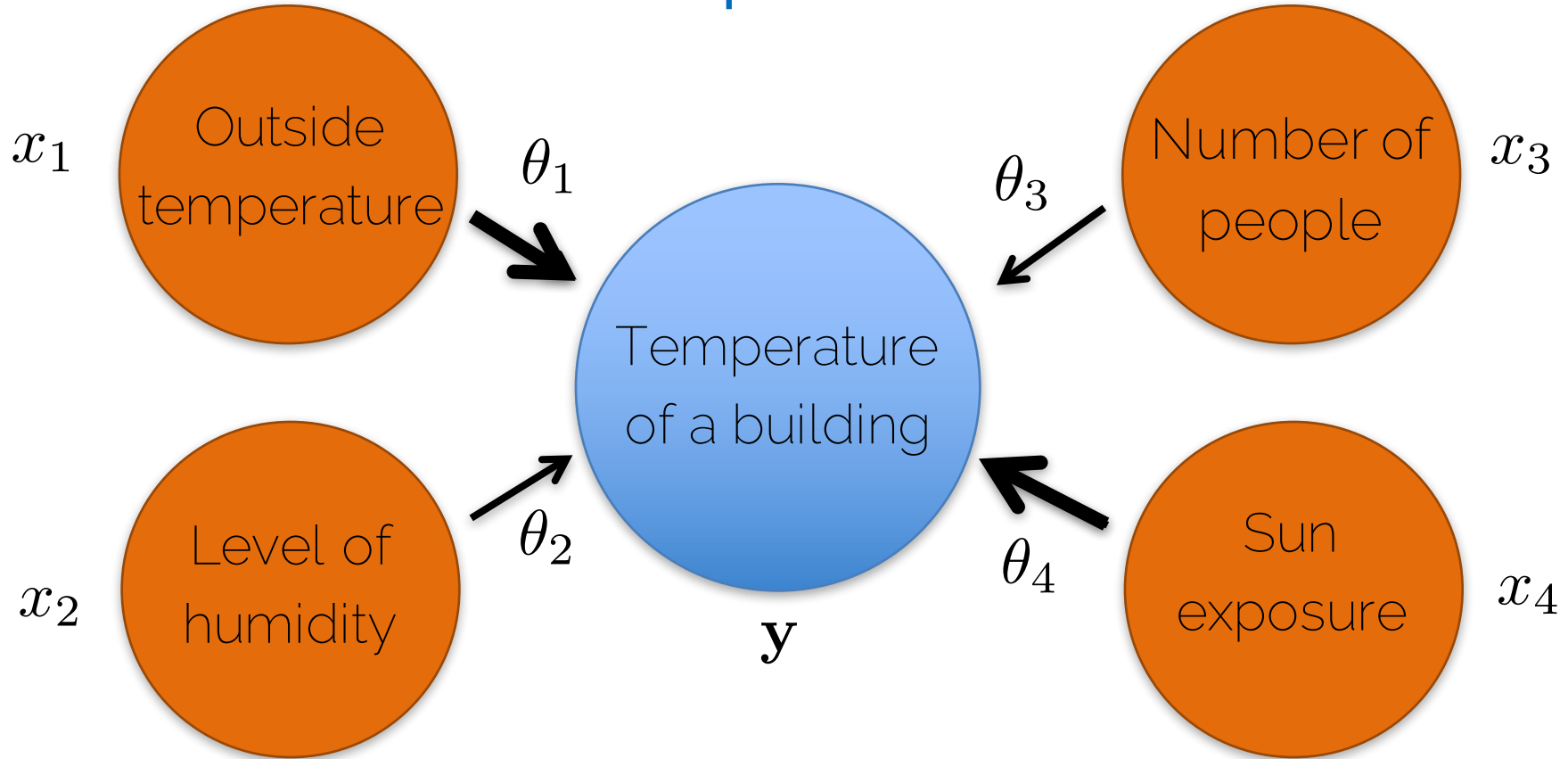
Linear prediction

- A linear model is expressed in the form

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



Linear prediction



Linear prediction

Prediction

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

Input features

Model parameters

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

Linear prediction

Temperature of the building

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

The diagram illustrates a linear prediction model. The input features are arranged in a matrix, with labels above each column: 'Outside temperature' (values 25, -10), 'Humidity' (values 50, 50), 'Number people' (values 2, 0), and 'Sun exposure (%)' (values 50, 10). The output is a vector of predicted temperatures \hat{y}_1 and \hat{y}_2 . The model coefficients are shown in a column vector on the right, with values 0.64, 0, 1, and 0.14. The coefficients 0, 1, and 0.14 are highlighted with orange boxes, and orange arrows point from the corresponding feature values (50, 0, 10) in the input matrix to these coefficients.

Linear prediction

How do we
obtain the
model?



Temperature
of the building

$$\begin{bmatrix} 25 \\ -5 \end{bmatrix}$$

=

Outside
temperature

$$\begin{bmatrix} 25 \\ -10 \end{bmatrix}$$

Humidity

$$\begin{bmatrix} 50 \\ 50 \end{bmatrix}$$

Number people

$$\begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

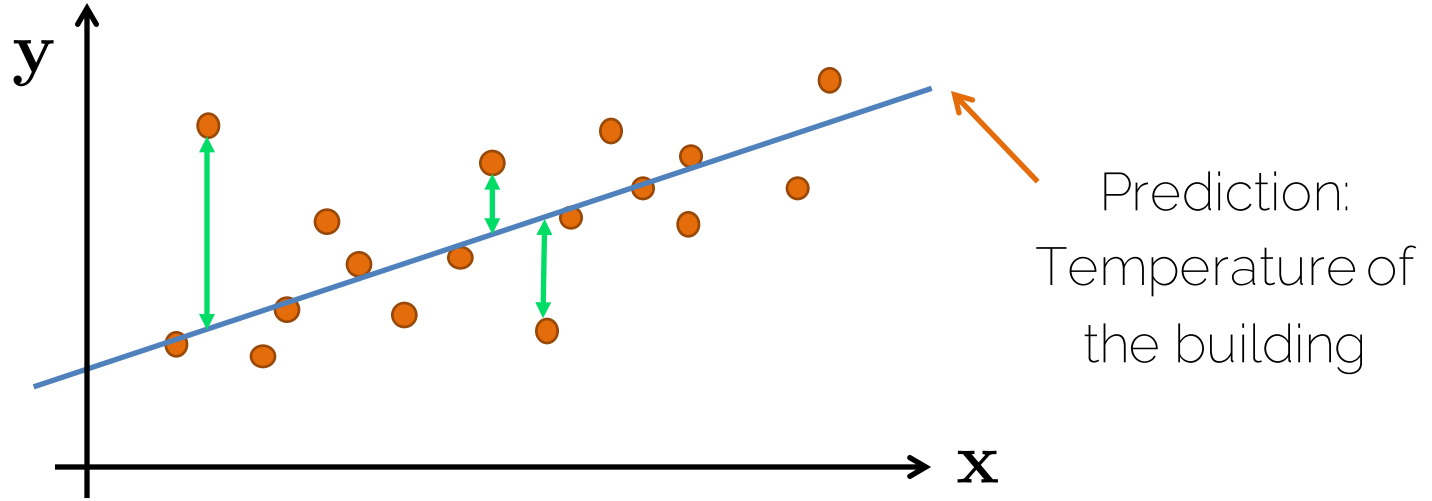
Sun exposure (%)

$$\begin{bmatrix} 50 \\ 10 \end{bmatrix}$$

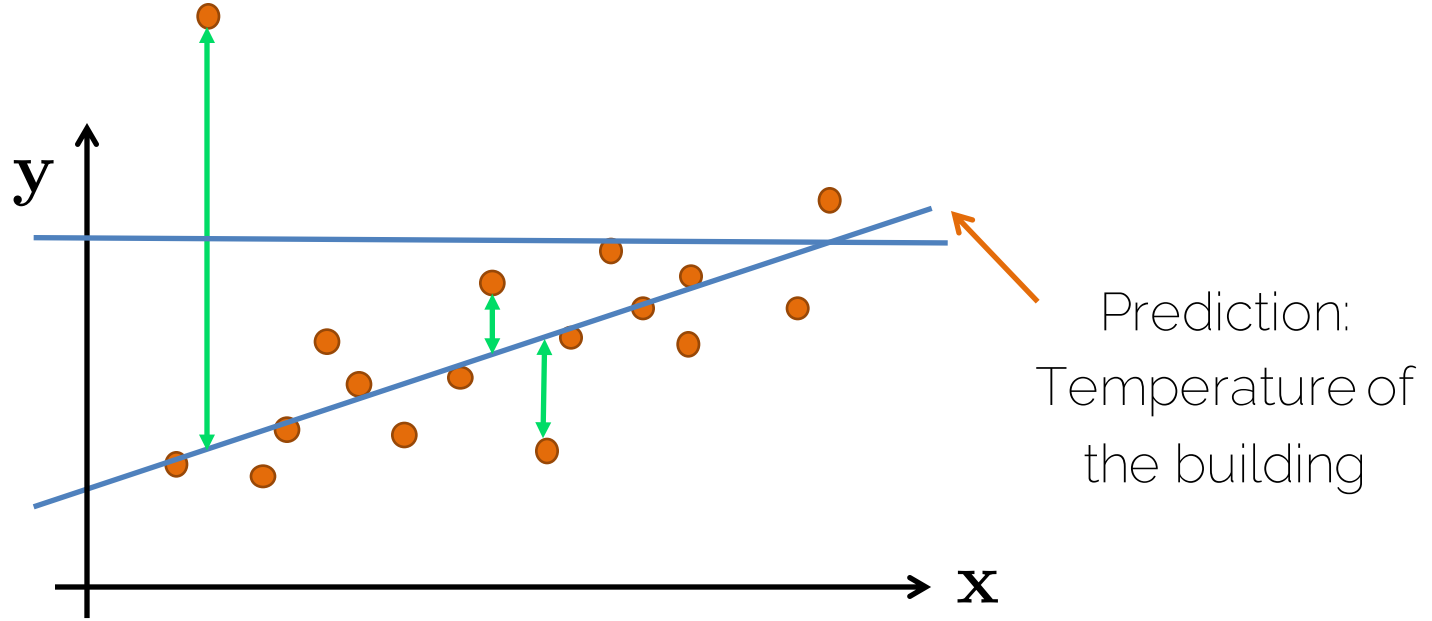
MODEL

$$\begin{bmatrix} 0.64 \\ 0 \\ 1 \\ 0.14 \end{bmatrix}$$

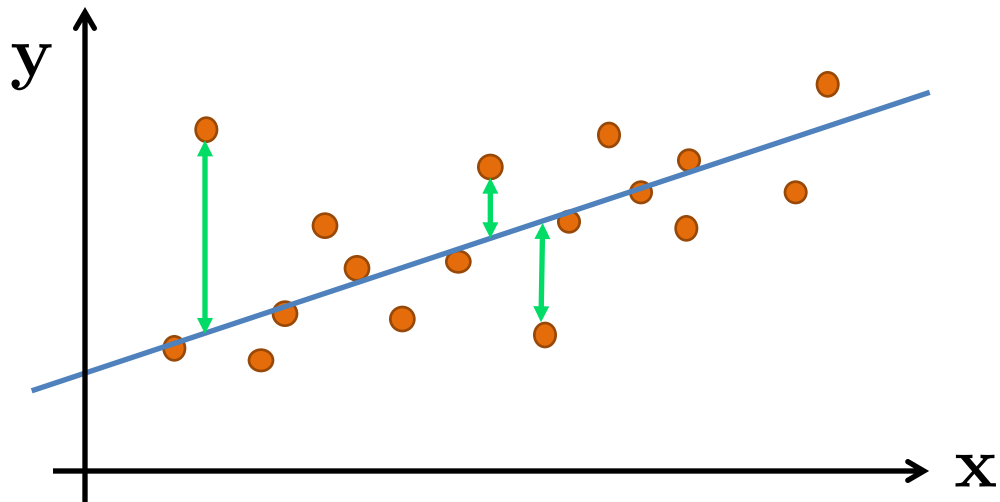
Linear regression



Linear regression



Linear regression



Minimizing

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

Objective function
Energy
Loss

Optimization

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

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



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

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?

- Mean squared error (MSE)

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

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.

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

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

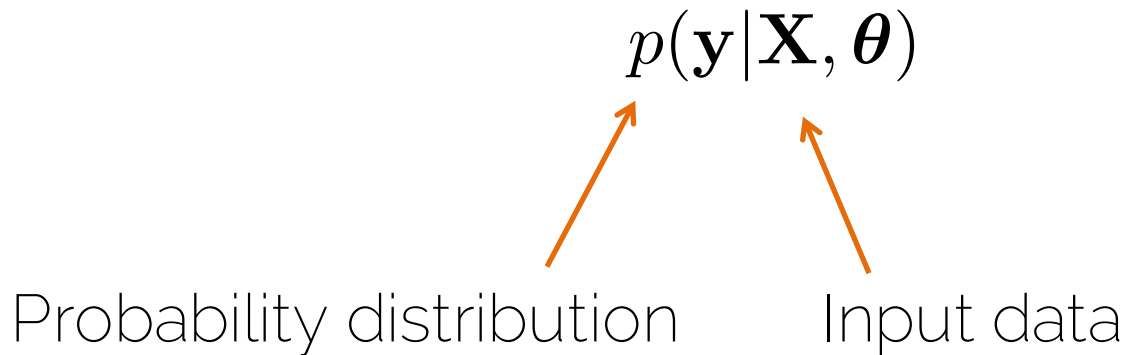
Maximum Likelihood Estimate

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

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

Spoiler: Related to softmax loss

Back to linear regression



Back to linear regression

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

i.i.d. = independent and
identically distributed



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

Back to linear regression

$$p(y_i | \mathbf{x}_i, \boldsymbol{\theta})$$

Gaussian or Normal
distribution

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

mean

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

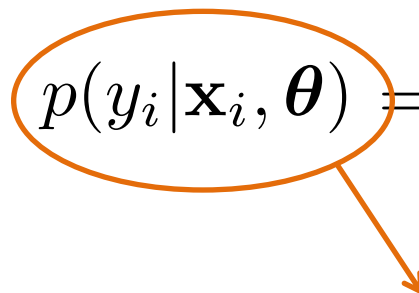
Back to linear regression

$$p(y_i | \mathbf{x}_i, \boldsymbol{\theta})$$


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


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

Back to linear regression

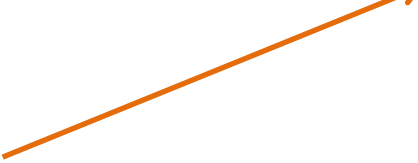
$$p(y_i | \mathbf{x}_i, \boldsymbol{\theta}) = (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2} (y_i - \mathbf{x}_i \boldsymbol{\theta})^2}$$


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




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


Back to linear regression

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


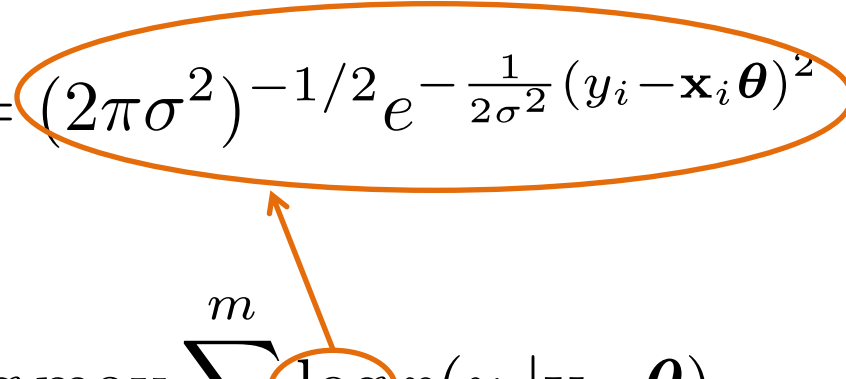
Assuming $y_i = \mathcal{N}(\mathbf{x}_i \boldsymbol{\theta}, \sigma^2) = \mathbf{x}_i \boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$



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

Back to linear regression

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

$$\boldsymbol{\theta}_{ML} = \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^m \log p(y_i | \mathbf{x}_i, \boldsymbol{\theta})$$
An orange oval highlights the term $\log p(y_i | \mathbf{x}_i, \boldsymbol{\theta})$ in the summation of the maximum likelihood estimation formula. An orange arrow points from this term up to the corresponding term in the likelihood function equation above.

Back to linear regression

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

Matrix notation


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

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

Back to linear regression

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

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


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

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

How can we
find the
estimate of
theta?

Back to linear regression

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

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

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

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

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

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

Regularization

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

L2 regularization

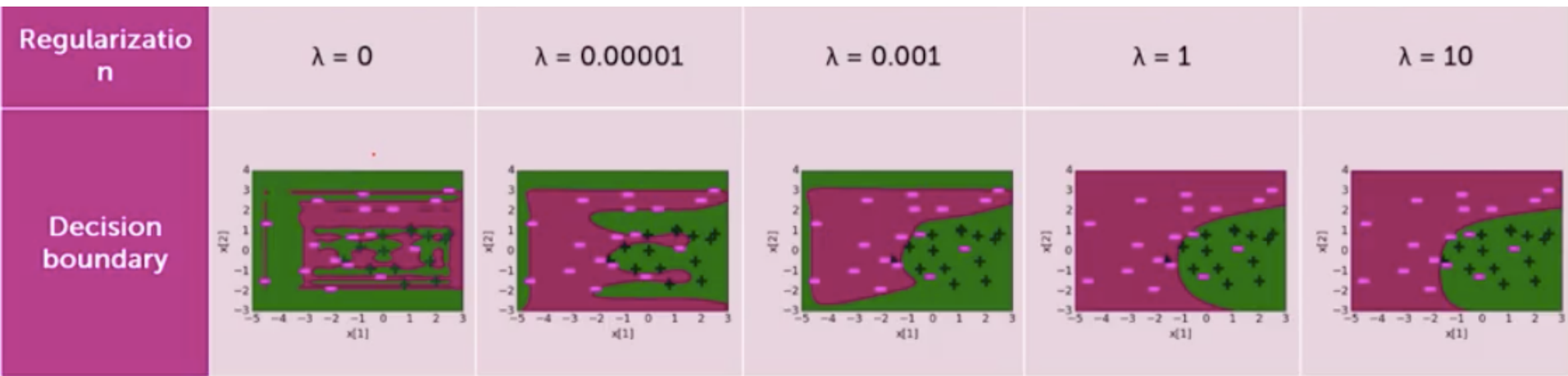
L1 regularization

Max norm regularization

Dropout

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

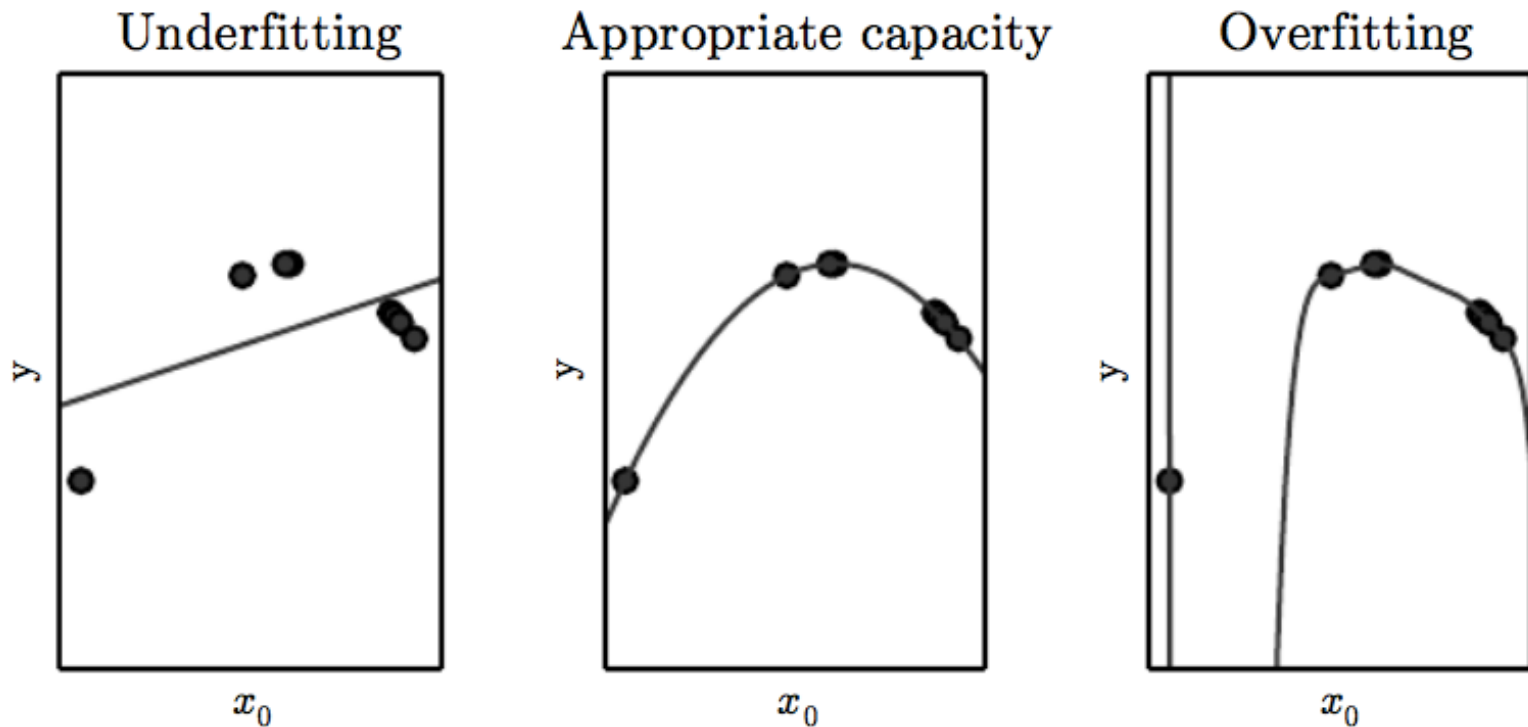
Regularization



What is the goal of regularization?

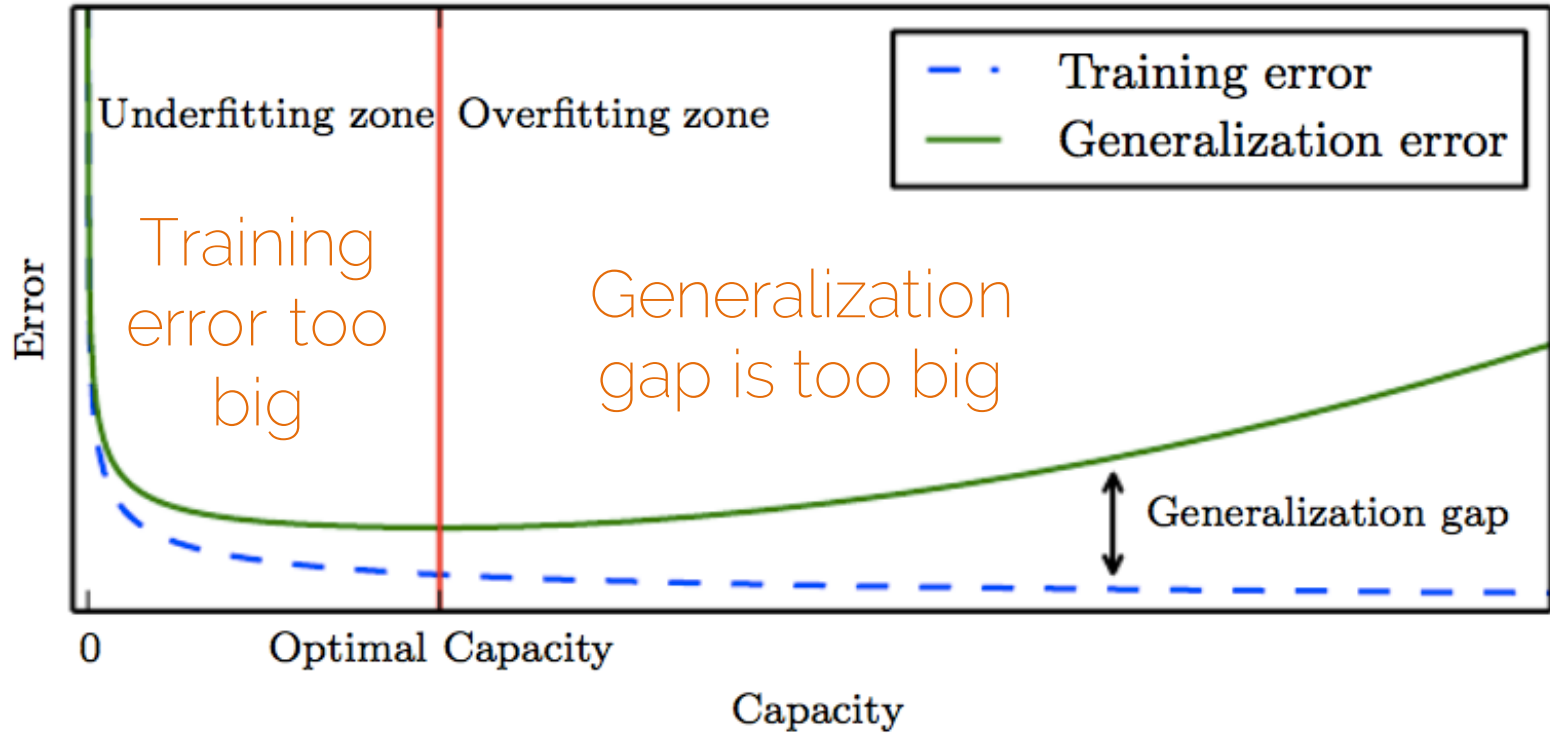
What happens to the training error?

Overfitting and underfitting



What is lambda for each of the cases?

Overfitting and underfitting



Live demo

Next lectures

- Thursday 4th May in the Chemistry Building!
- Topic: optimization
- First exercise on Friday 5th of May here!