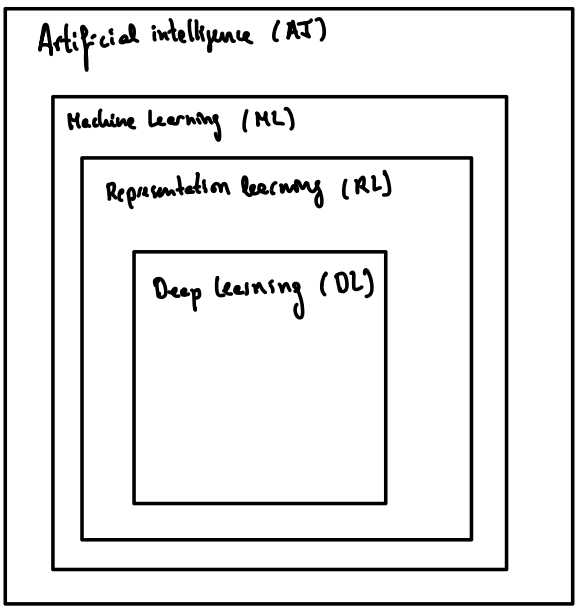
② Machine learning: Introduction

* Machine learning / Artificial intelligence / Deep learning / Representation learning



AI: { Knowledge Base
Example: Infering after some statements/logics
"Reasoning"

ML: { extracting patterns from row data
Example: logistic regression

RL: { Finding convinient ways of representing the data (features)
Example: PCA / autoencoders

DL: { · Representations expressed in terms of simpler representations (large number of unknowns)
· Example: ANN

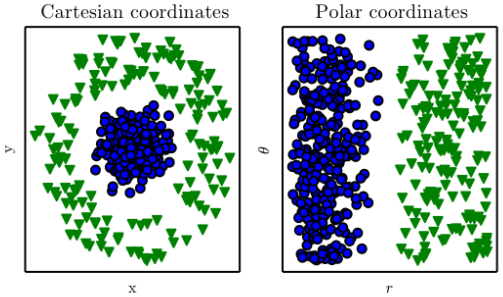
* Supervised / Unsupervised learning

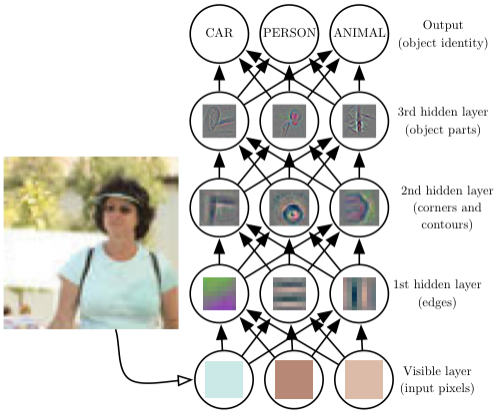
* Supervised : $D = \{x^i, y^i\}$ (features, outputs)
  * Example: y = temperature; x_1 = day; x_2 = place;
  * Cases: regression (linear, polynomial); classification
* Unsupervised : $D = \{x^i\}$
  * Cases: Clustering; Dimension reduction

* Predictions - Representation matters

* $y_i \simeq f_\theta(x_i)$;
* Linear predictor: $f_\theta(x_i) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$
* Polynomial predictor (Machine learning): $f_\theta(x_i) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1 x_2 + \theta_5$
* Representation learning example: $f_\theta(x_i) = \theta_i f_\theta^i(x)$ (learn also the representation)

  ¿ w? ¿ exp?
* Deep-learning example: $f_\theta(x_i) = \theta_i f_\theta^i(f_\theta(f_\theta(\dots(x))\dots))$





* **First example**: Linear regression in l^2 norm

$\hat{y} = X\theta$; $X = \begin{bmatrix} 1, x_1, x_n \end{bmatrix}$ { all dat samples ; $y = \begin{bmatrix} \vdots \\ \vdots \end{bmatrix}$ outputs

$\theta = \{\theta_1, \theta_n, \theta_j\}$

$\min_{\theta} f(\theta) = MSE(\theta) = \| y - X\theta \|_2^2$ $\quad A = [\ ]$

- Normal equations: $\nabla f(\theta) = 0 \Rightarrow X^T X \theta = X^T y$
- Gradient method: $\theta_{k+1} = \theta_k - \varepsilon X^T [X\theta_k - y]$
- Geometric interpretation: $\min_{\theta, u} \| y - u \|_2^2$ $\quad u = X\theta$
  (Projection in the linear space generated by columns of A)

* **Second example**: Linear regression in l^1 norm

$\min_{\theta} f(\theta) = | y - X\theta |_1 = \sum_i^N | y_i - x_i \theta |$

- Convex problem / not differentiable
- Adding auxiliar variables (double of variables

$$\begin{cases} \min_{\theta, t} & 1^T t \\ & -t \leq y - X\theta \leq t \end{cases}$$

- Canonical form

$$\begin{cases} \min_{r} & c^T r \\ & Ar \leq b \end{cases}$$

with $\begin{cases} r = [\theta, t]^t, \ A = \begin{bmatrix} X & -I \\ -X & -I \end{bmatrix} \\ c^T = [0, 1^T]; \ b = \begin{bmatrix} y \\ -y \end{bmatrix} \end{cases}$

* **Third example**: Linear regression in l^∞ norm (Chebyshev approx)

$$\min_\theta f(\theta) = |y - X\theta|_\infty = \max_i |y_i - x_i \theta|$$

- Convex problem / not differentiable
- Adding auxiliar variables (adding 1 variables)

$$\begin{cases} \min_{\theta, t} \; t \\ -t\mathbf{1} \leq y - X\theta \leq t\mathbf{1} \end{cases}$$

- Canonical form

$$\begin{cases} \min_r \; c^T r \\ Ar \leq b \end{cases}$$

with $r = [\theta, t]^t$, $A = \begin{bmatrix} X & 1 \\ -X & -1 \end{bmatrix}$, $c^T = [0, 1]$, $b = \begin{bmatrix} y \\ -y \end{bmatrix}$

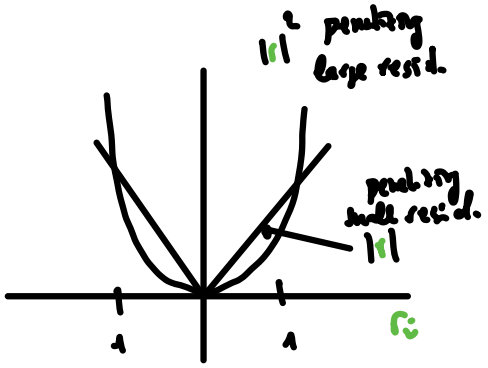
* Different norms: interpretation in terms of penalty function

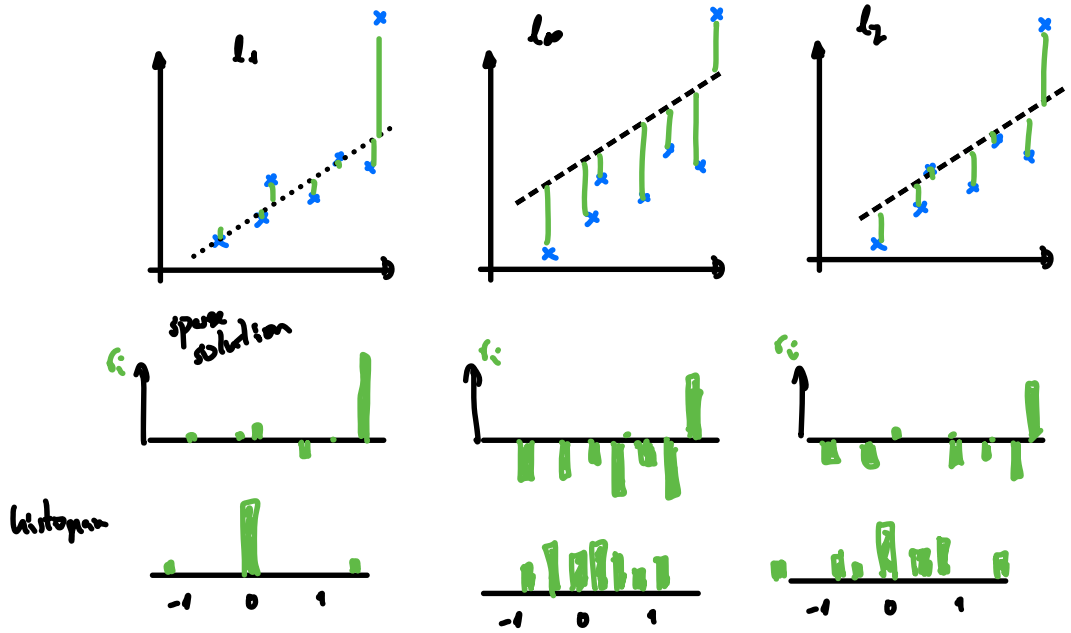
Defining $r_i = y_i - x_i \theta$

L_2: $\left[\sum \phi(r_i)\right]^2 = \sum r_i^2$

L_1: $\sum \phi(r_i) = \sum |r_i|$

L_∞: $\max_i |r_i|$





Other penalties:

$$\begin{cases} \phi(r) = r^p \\ \phi(r) = \begin{cases} -a^2 \log\left(1-\left(\frac{r}{a}\right)^2\right) & |r|<a \\ \infty & |r|\geq a \end{cases} \\ \phi(r) = \begin{cases} 0 & |r|<a \\ |r|-a & |r|\geq a \end{cases} \end{cases}$$

* Hypothesis space / Capacity / Overfitting

  * Why not polynomial regression? Hypothesis space:

$$\hat{y}_i = \theta_j x_i^j = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2 + \dots$$

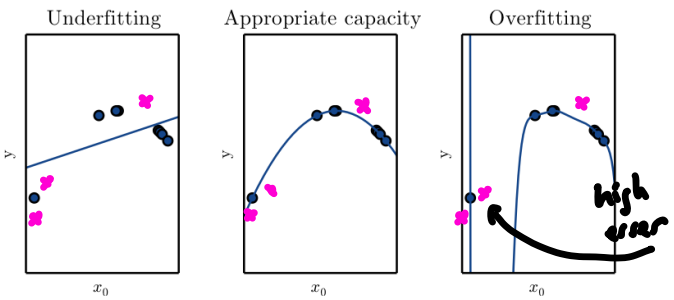
* ¿ Which dimension of polynomial?
  Is always useful to increase the dimension / capacity?

* Divide data in train (80%) and test (20%)

  - Train : to minimize MSE and obtain θ (MSE^{train})
  - Test : validate (MSE^{test})

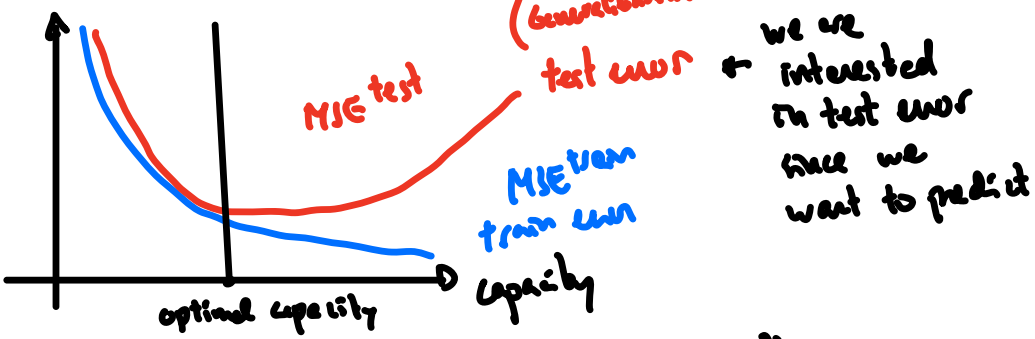
* Optimization vs ML
  - Optimization goal: find best minimizers
  - ML goal: predict

* Capacity



- Too low capacity: **Underfitting**
  - High error in train
  - Too few variables θ in optim.
  - Test error also high
- Appropriate capacity:
  - low train error
  - low test error
- High capacity: **Overfitting**
  - Very low train error
  - High test error

* Train/test error vs capacity

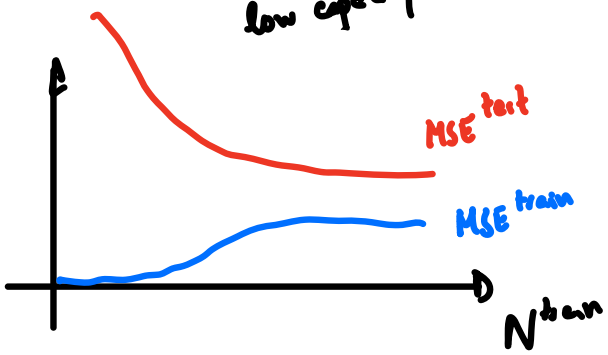


we are interested in test error since we want to predict

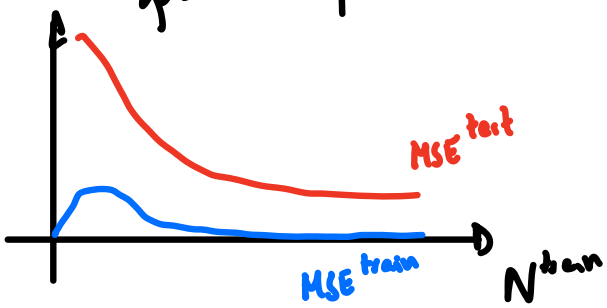
$$\| MSE^{test} - MSE^{train} \|^2 \leq f(c, N^{\#train})$$

capacity, number of data

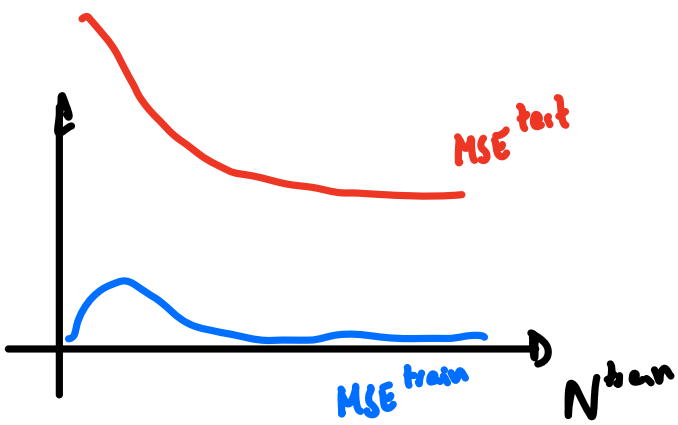
low capacity



optimal capacity



high capacity



* Underfitting and overfitting from the optimization point of view

$$\min \|y - X\theta\|^2$$

$$(X^TX)\theta = X^Ty$$

Underfitting: $X = \begin{bmatrix} & \\ & \end{bmatrix}$;

$X^TX = \begin{bmatrix} & \\ & \end{bmatrix}$; θ is too small in comparison with data

Overfitting: $X = \begin{bmatrix} & & \\ & & \end{bmatrix}$; $X^TX = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}$

X^TX not full rank

θ is too large in comparison with data

* Regularization

* L^2 regularization (Tikonov regularization / Ridge regression)

$$\min \|y - X\theta\|_2^2 + \lambda \underbrace{\|\theta\|^2}_{\theta^T\theta}$$

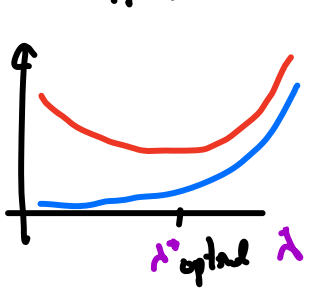
- $\lambda > 0$
- Prior knowledge of θ
- Penalize large values of θ
- Also useful when X is square but bad conditioned (Elasticity?)

$$\underbrace{(X^TX + \lambda I)}_{\text{full rank}}\theta = X^Ty$$

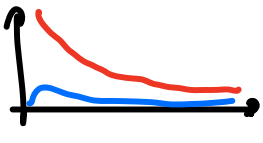
* Other possible regularization
  - L^1 regularization $|\theta|$
  - $\|\nabla\theta\|^2 \rightarrow$ smoothing (Laplacian)
  - $|\nabla\theta| \rightarrow$ Total variation

* Effect of regularization

. Possible to use large capacity with appropriate λ value



λ and dimensions are called hyperparameters



with optimal λ

* Cross-validation

* Useful when data is limited
* Divide data $D = \bigcup D_i$ with $D_i \cap D_j = 0$ (around 5/6 times)
* Compute mean and standard deviation

$$\mu = \frac{1}{L} \sum_i^L MSE_i^{test}$$

$$\sigma^2 = \frac{1}{L} \sum_i^L (MSE_i^{test} - \mu)^2$$

we plot $\{\mu \pm 1.96\,\sigma\}$