Parameters vs hyperparameters

CEMS
Chemical Engineering & Materials Science

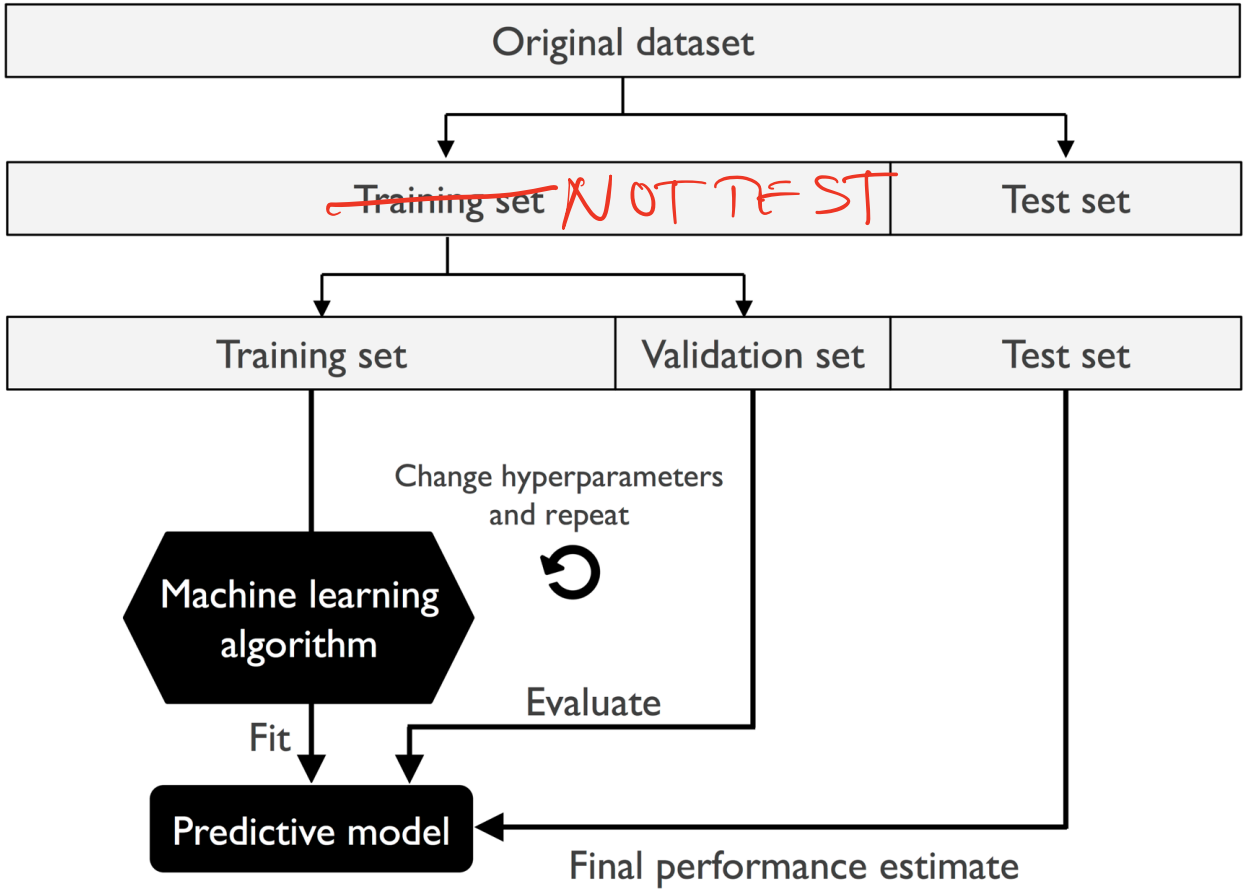
$$\hat{y} = f(\vec{x}) = 3x_1^2 + 2x_2$$

parameters = learned during training to minimize loss
(weights)
(coefficients)

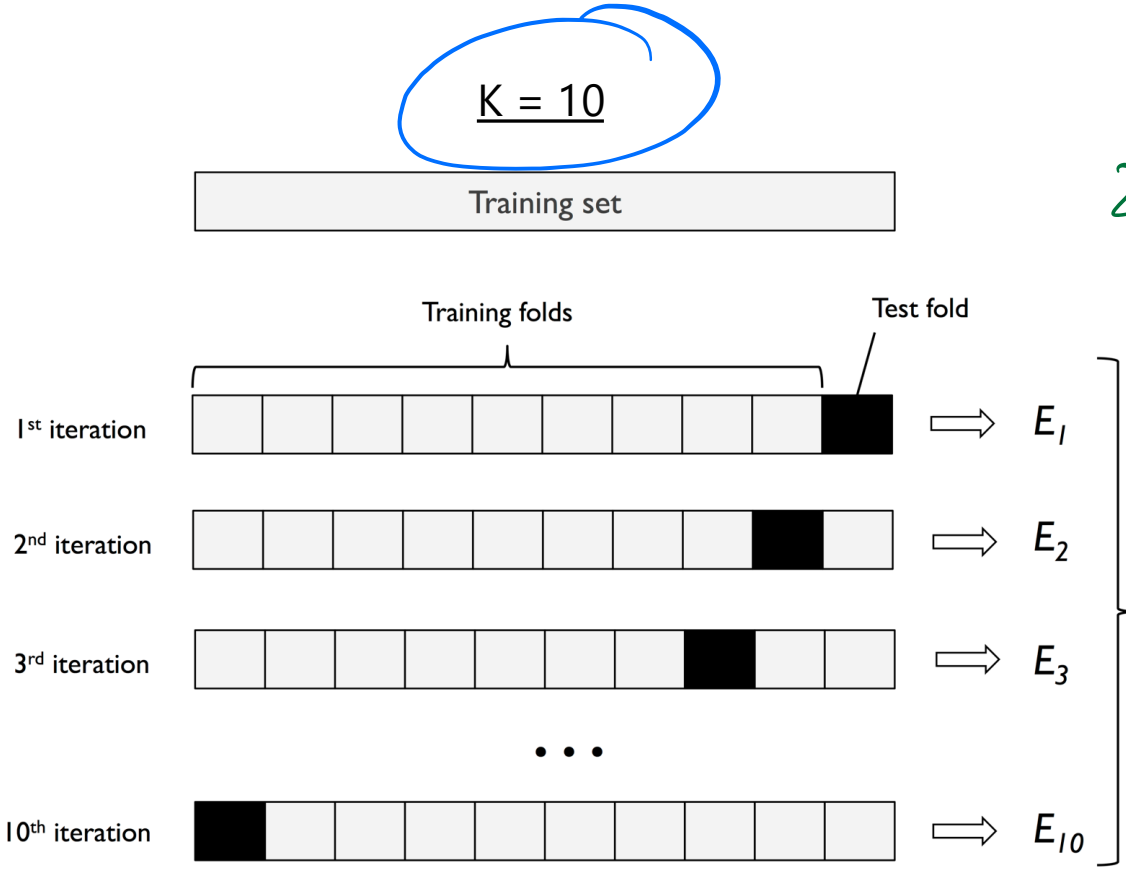
hyperparameters = knobs that were turned to arrive at $f(\vec{x})$

- ↳ # features
- ↳ loss function
- ↳ model type

Validation and testing workflow



K-fold cross validation



1) Choose hyperparameters

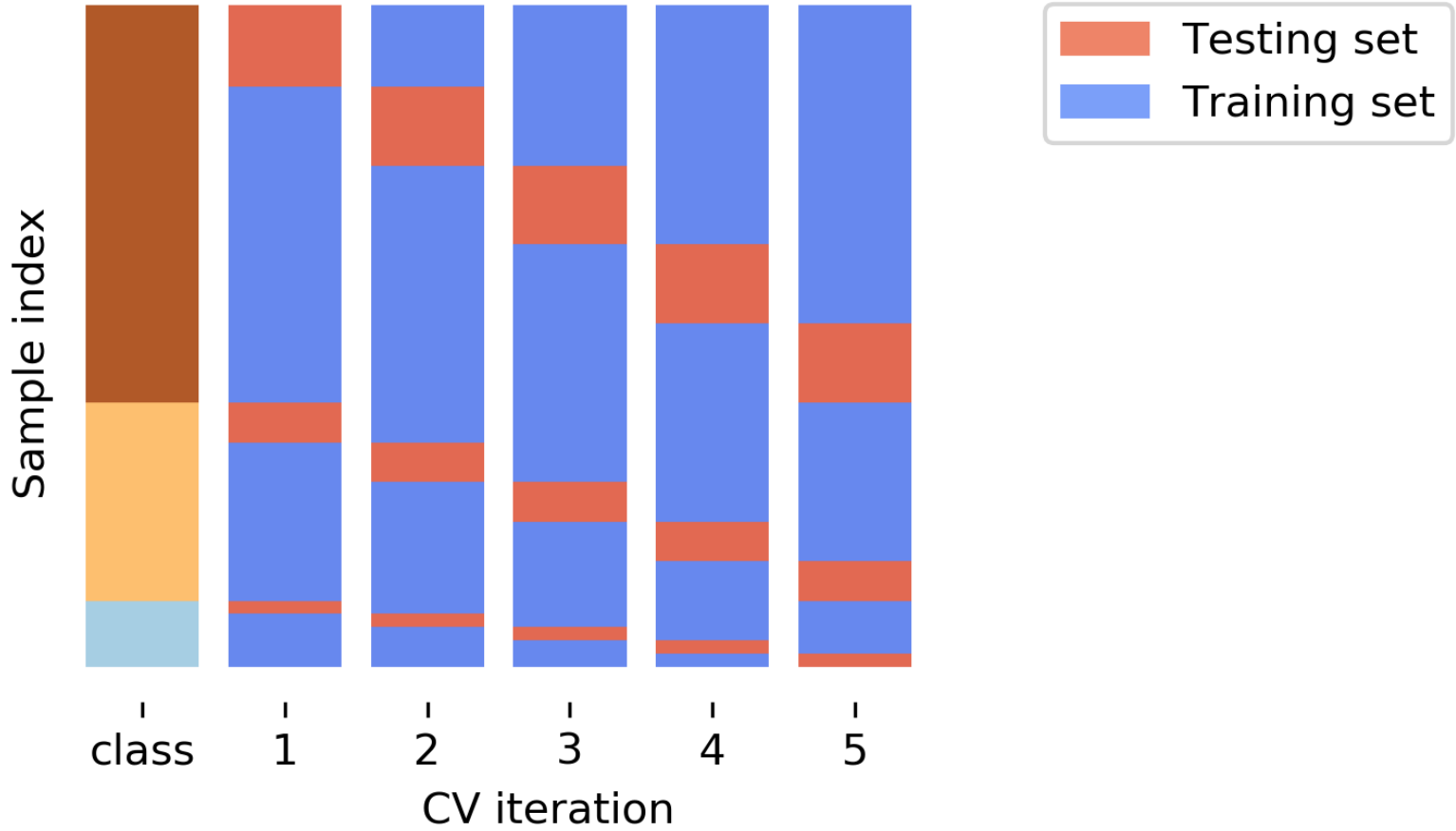
2) Run CV

3) Average scores

↑ K ⇒ more robust estimate of validation performance

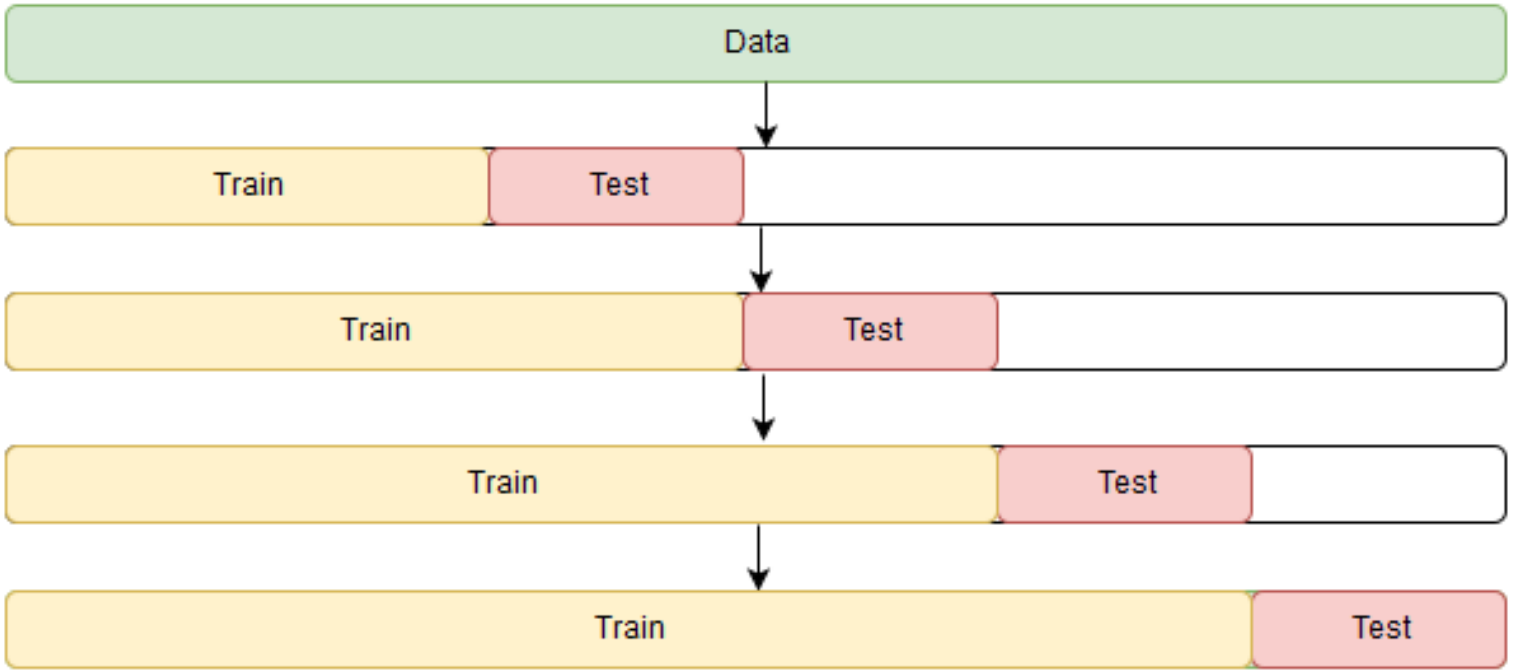
↑ K ⇒ ↑ cost

Class imbalance: stratified cross validation



Time series: "Rolling validation"

$t_0, t_1, t_2, \ldots$, t_f.



Understanding data validation: IID vs OOD

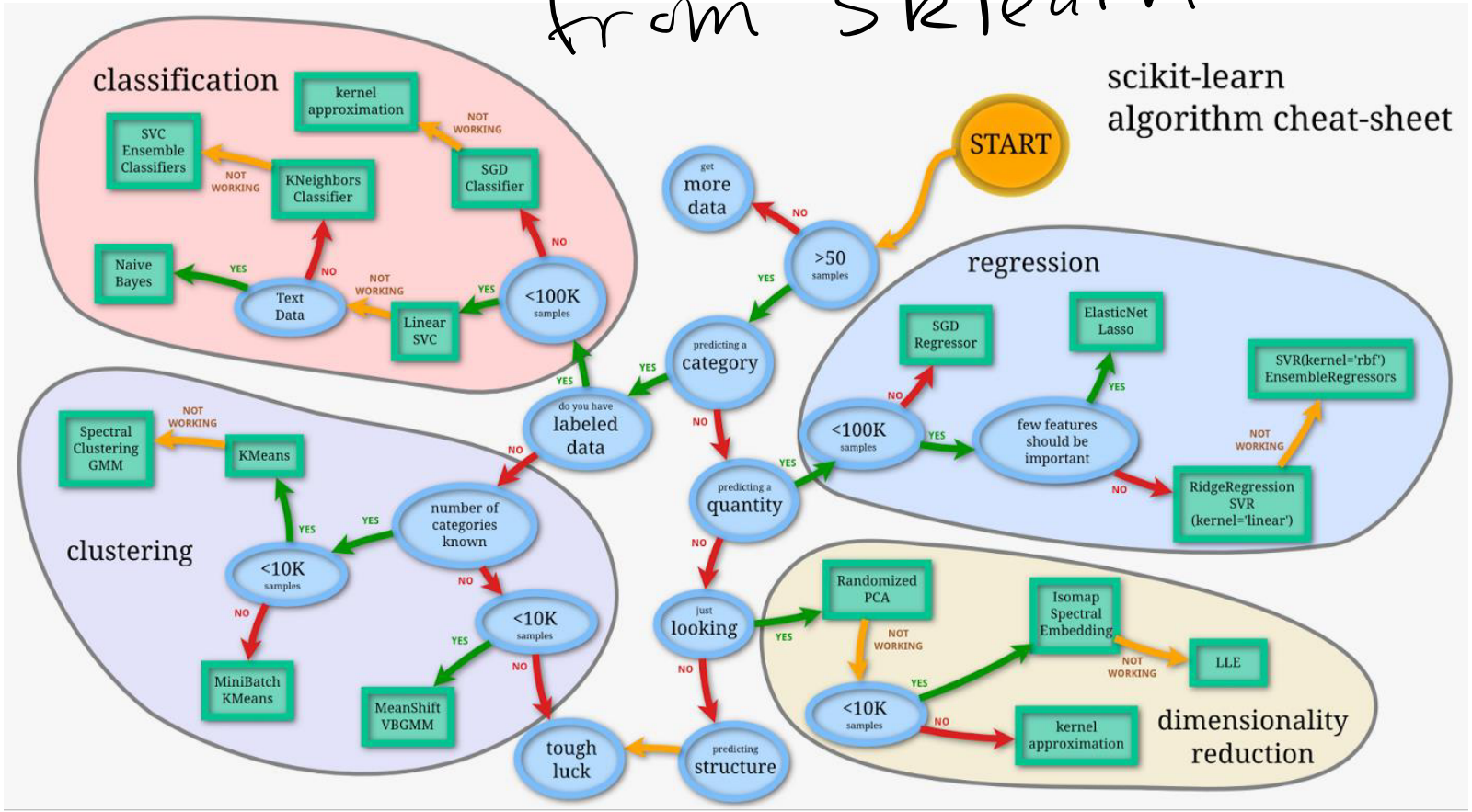
IID ⇒ independent, identical distribution
(sample randomly from same data as training)

OOD ⇒ out of distribution
(sampling points that are distinct from training data)

ML models ⇒ designed for IID
⇒ OOD sometimes correlated w/ IID

Which model to use??

from sklearn



Basics of linear regression

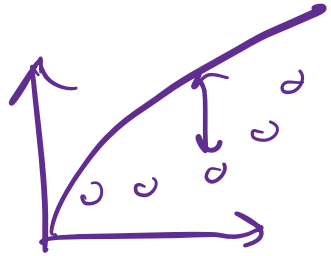
$$\hat{y} = f(\vec{x}) = \vec{\omega} \cdot \vec{x} = \omega_0 \vec{x}_0 + \omega_1 \vec{x}_1 + \omega_2 \vec{x}_2$$

Training → find $\vec{\omega}$ that minimize loss, $\mathcal{L}(y, \hat{y})$

Loss functions

$$\text{OLS (MSE)} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \vec{\omega} \vec{x}_i)^2$$

$$\vec{\omega} : \min_{\vec{\omega}} \| \vec{X} \vec{\omega} - y \|_2^2$$



OLS ⇒ can be solved analytically (linear algebra)

loss = error

Basics of linear regression

Regularized regression

$$\text{LASSO}\ (L_1) = \text{MSE} + \alpha \|\vec{w}\|_1 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{K} |w_j|$$

$K \Rightarrow$ # features

$\alpha \Rightarrow$ regularization parameter

$$\text{Ridge}\ (L_2) = \text{MSE} + \alpha \|\vec{w}\|_2^2 = \frac{1}{N}\sum(y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{K} w_j^2$$

Discourages overfitting.
by penalizing large weights, w_j

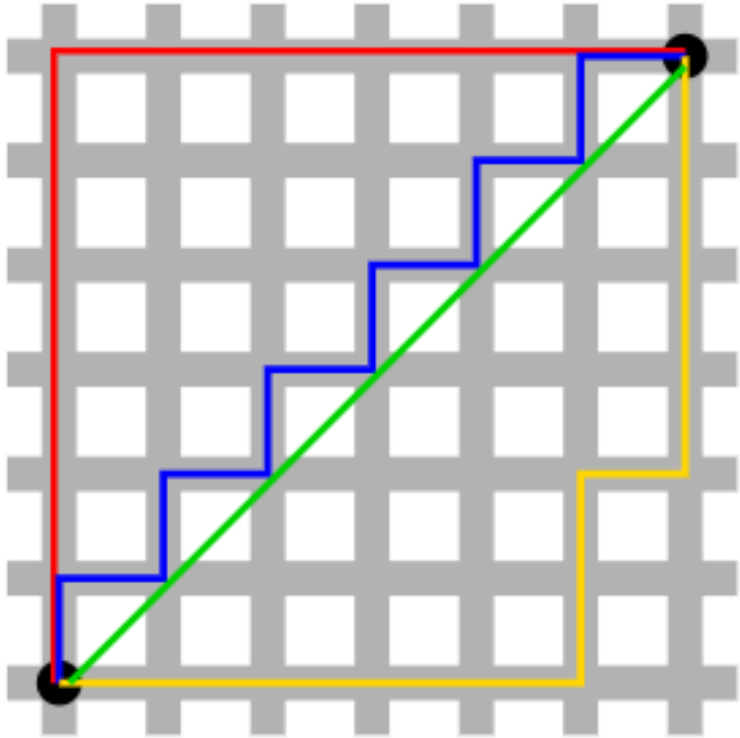
Can't be solved analytically, so need training.

L1 vs L2 norm

$$\hat{y}(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p$$

LASSO (L1 norm): $\min_w ||Xw - y||_2^2 + \alpha ||w||_1$

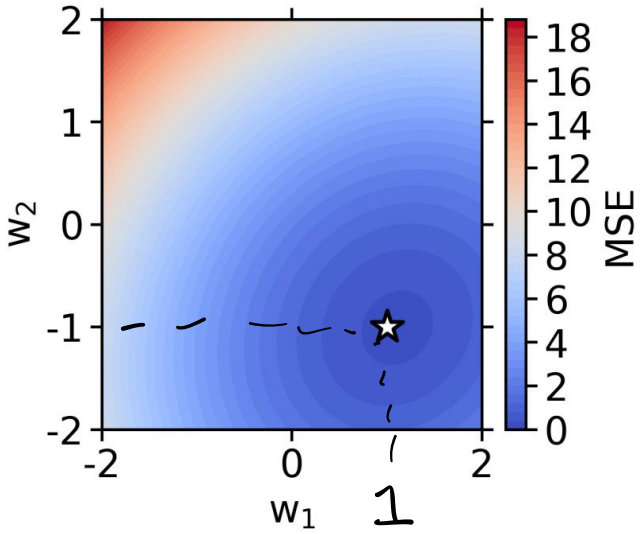
Ridge (L2 norm): $\min_w ||Xw - y||_2^2 + \alpha ||w||_2^2$

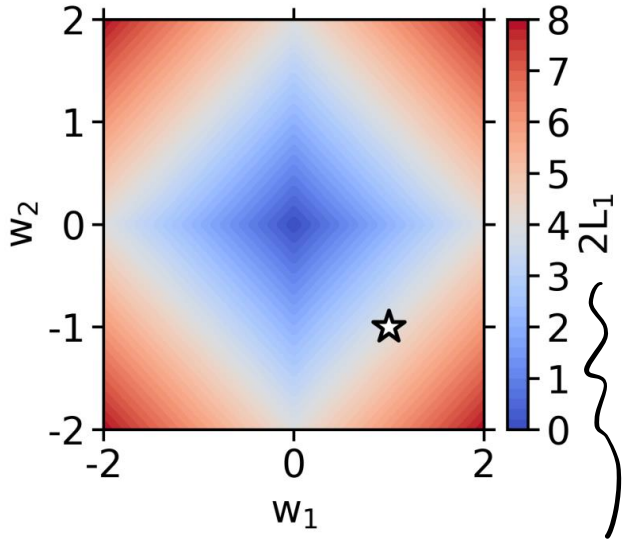


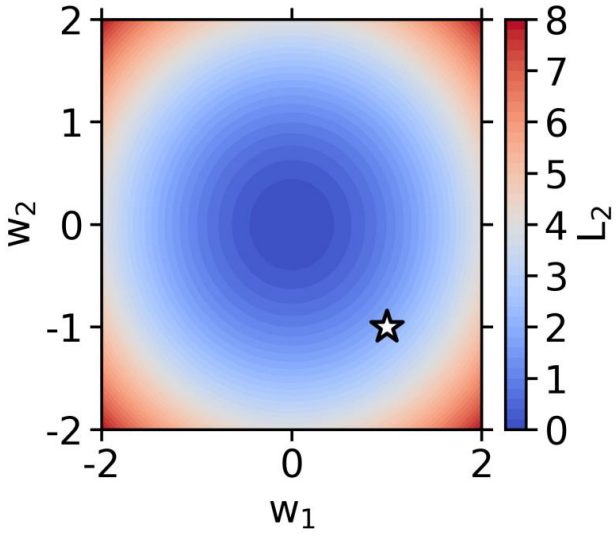
L1 vs L2 norm

TRUE → $y = 1.0x_1 - 1.0x_2 + N(0, 0.25)$

$\hat{y} = w_1 x_1 + w_2 x_2$







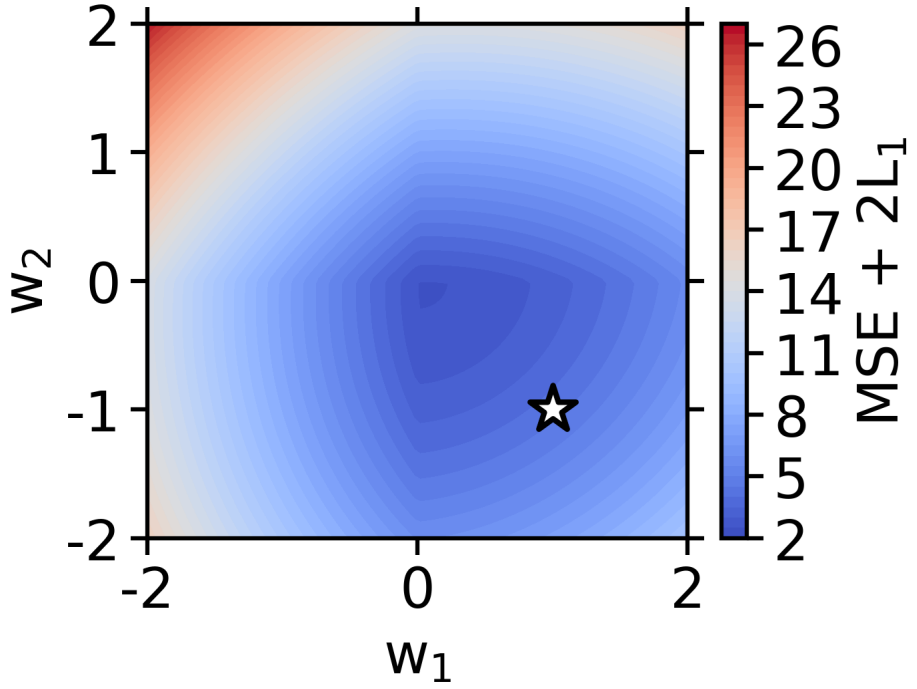
$2L_1 = 2(|w_1| + |w_2|)$

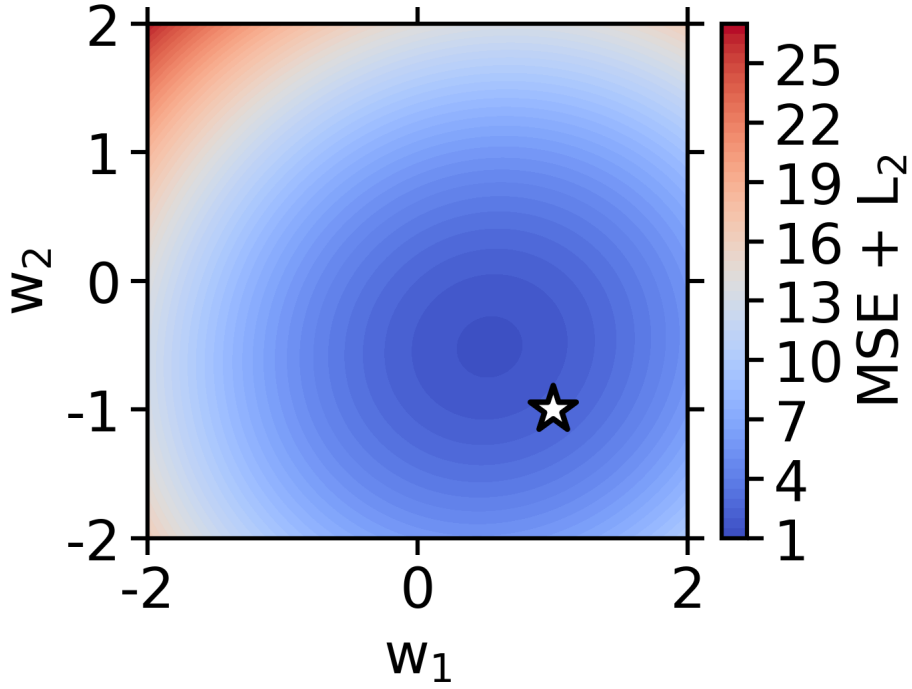
$L_2 = w_1^2 + w_2^2$

L1 vs L2 norm

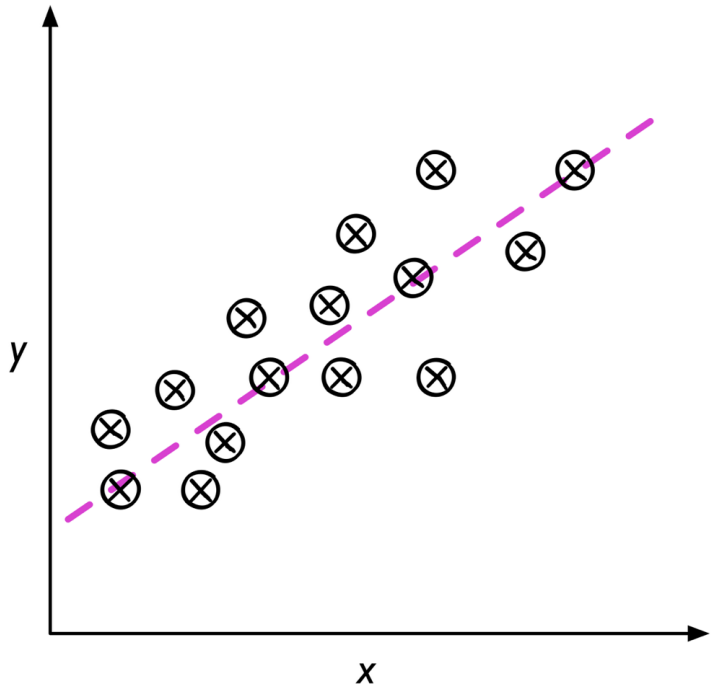
$$y = 1.0x_1 - 1.0x_2 + N(0, 0.25)$$

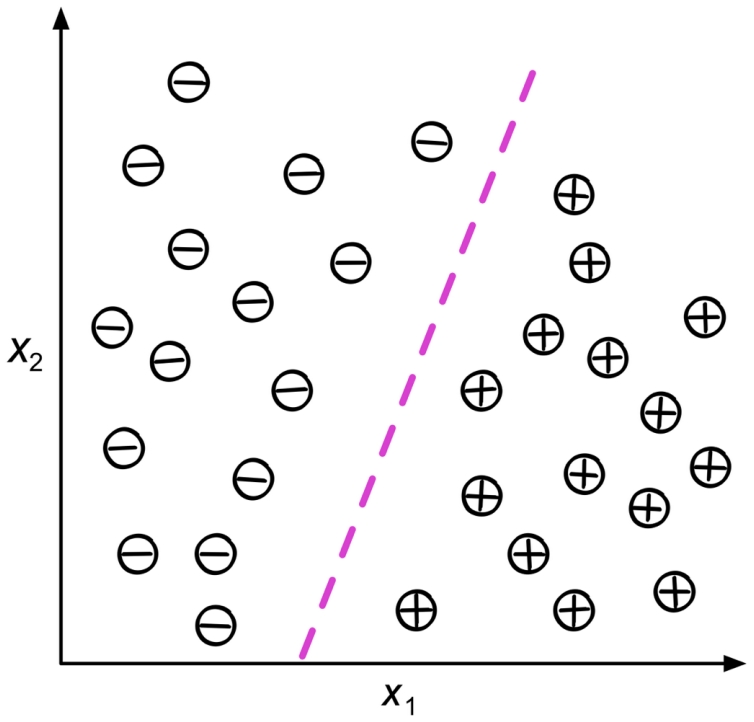
$$\hat{y} = w_1 x_1 + w_2 x_2$$





How does this look for classification?





How does this look for classification?

For binary classification, y is no longer continuous, but binomial:

y = [1, 1, 1, -1, -1, 1, -1, -1, …]