ELTE | FACULTY OF INFORMATICS

REGULARIZATION

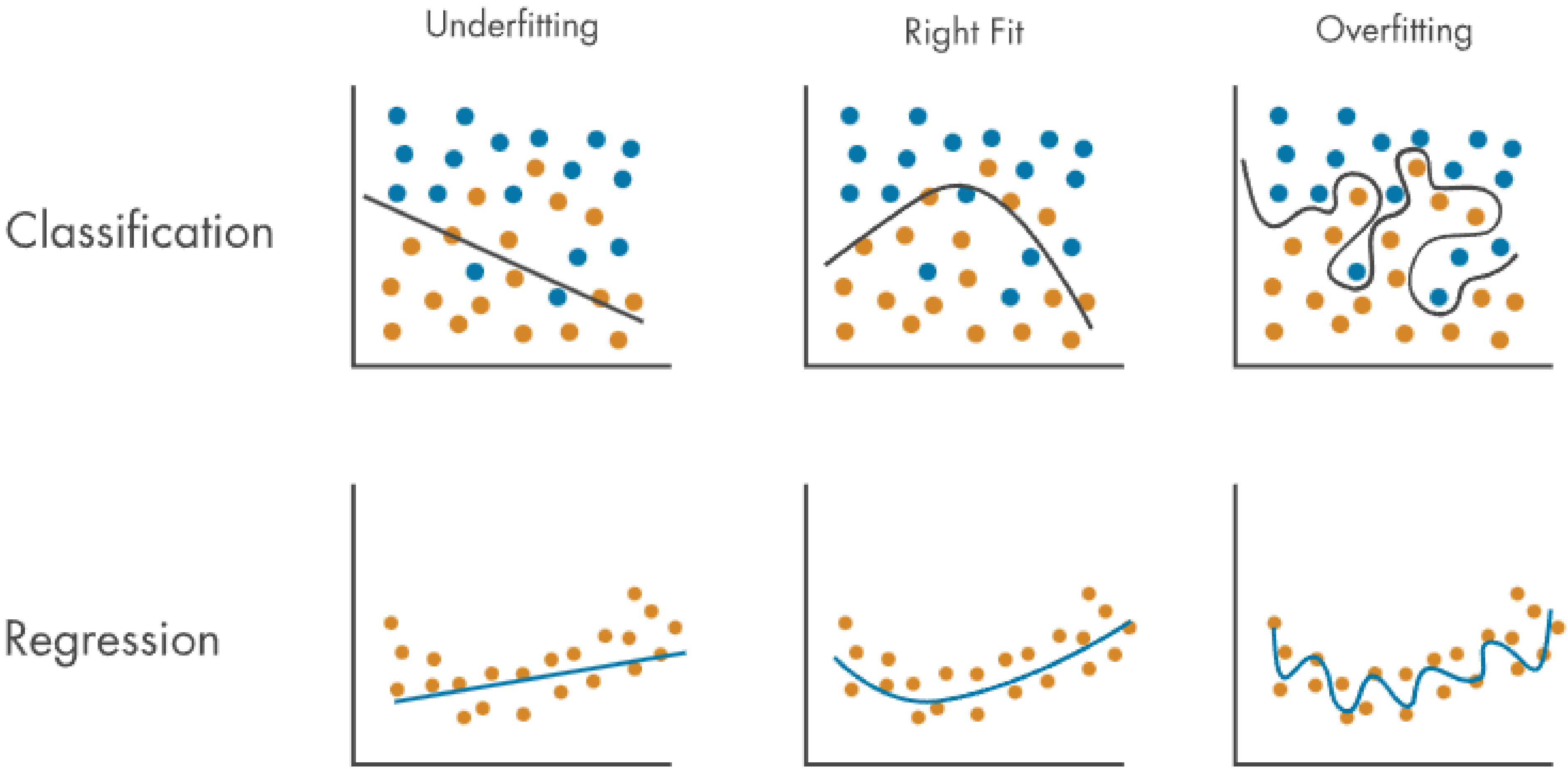
Machine Learning Course
Balázs Nagy, PhD

ELTE | IK
DEPARTMENT OF ARTIFICIAL INTELLIGENCE

Fit problems

High Bias

High Variance



Fit problems

- **Underfit:**
  - Model is too simple, need more features
- **Right fit:**
  - Nothing to do, model is good
- **Overfitting**:
  - If we have too many features, the learned hypothesis may fit the training set very well, but fail to generalize to new examples (predict on new examples)
  - Model is too complex, has too many features

Fit problems

- **Underfit:**
  - Model is too simple, need more features
- **Right fit:**
  - Nothing to do, model is good
- **Overfitting**:
  - If we have too many features, the learned hypothesis may fit the training set very well, but fail to generalize to new examples (predict on new examples)
  - Model is too complex, has too many features

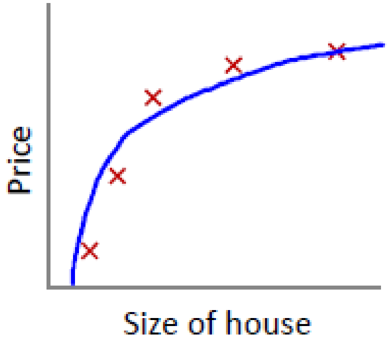
How to prevent overfit?

Prevent overfitting

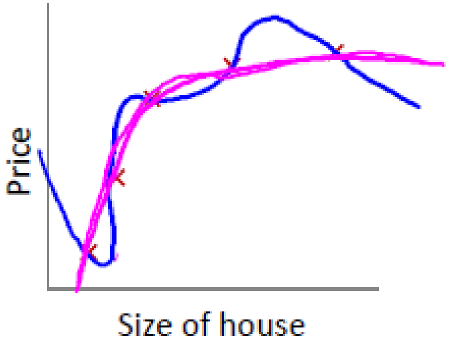
- Reduce number of features
  - Manually select which features to keep
  - Model selection algorithm
- Regularization
  - Keep all the features, but reduce magnitude / values of parameters Θ
  - Works well when we have a lot of features, each of which contributes a bit to predicting y

Regularization – Linear Regression

- Suppose we penalize high rank element and make w_3, w_4 really small



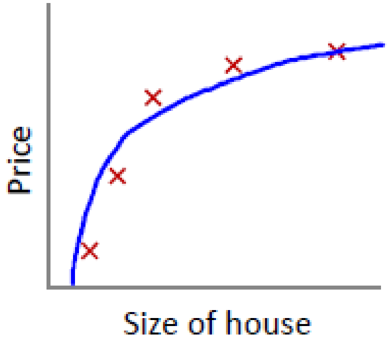
$$w_0 + w_1 x + w_2 x^2$$



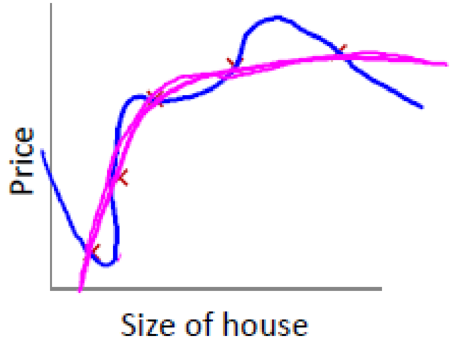
$$w_0 + w_1 x + w_2 x^2 + \cancel{w_3} x^3 + \cancel{w_4} x^4$$

Regularization – Linear Regression

- Suppose we penalize high rank element and make w_3, w_4 really small



$$w_0 + w_1 x + w_2 x^2$$



$$w_0 + w_1 x + w_2 x^2 + \cancel{w_3} x^3 + \cancel{w_4} x^4$$

$$\max_w \frac{1}{2m} \sum_{i=1}^{m} (h_w(x^{(i)} - y^{(i)})^2 + \lambda \sum_{j=1}^{n} w_j^2$$

NOTE: w_0 is **not** penalized

- The λ is the regularization parameter

Regularized Gradient descent

Repeat {

$$w_0 := w_0 - \mu \frac{1}{m} \sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) \cdot x_0^{(i)})$$

$$w_j := w_j - \mu \left[\left(\frac{1}{m} \sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) \cdot x_0^{(i)} \right) + \frac{\lambda}{m} w_j^2 \right] \qquad j \in 1, 2, \ldots, n$$

}

Regularized Gradient descent

Repeat {

$$w_0 := w_0 - \mu \frac{1}{m} \sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) \cdot x_0^{(i)})$$

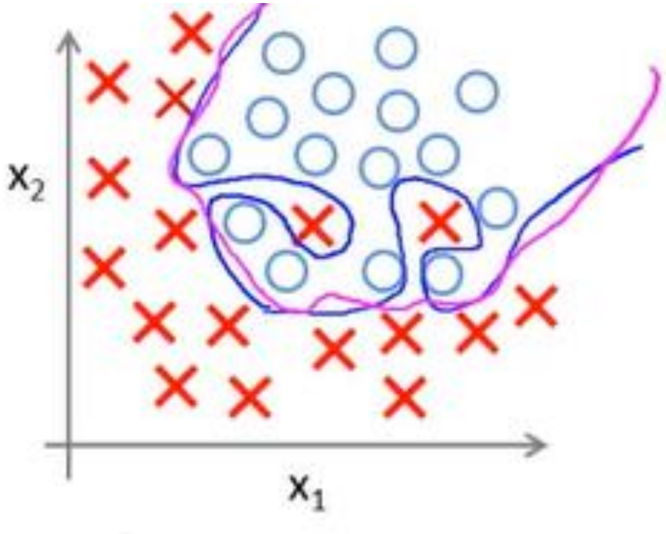
$$w_j := w_j - \mu \left[\left(\frac{1}{m} \sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)}) \cdot x_0^{(i)} \right) + \frac{\lambda}{m} \boxed{w_j^2} \right] \qquad j \in 1, 2, ..., n$$

}

L1 regularization: $|w_j|$

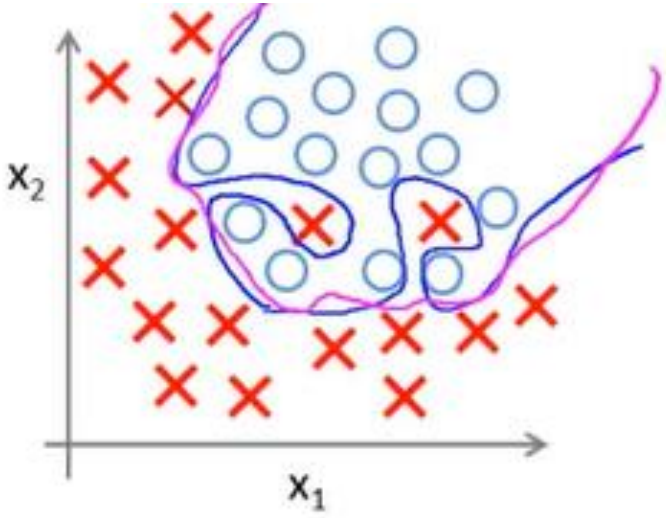
L2 regularization: w_j^2

Regularization – Logistic Regression



$$h_w(x) = g(w_0 + w_1 x_1 + w_2 x_2 +$$
$$w_3 x_1^2 + w_4 x_1^2 x_2 +$$
$$w_5 x_1^2 x_2^2 + w_6 x_1^3 x_2 + \ldots)$$

Regularization – Logistic Regression



$$h_w(x) = g(w_0 + w_1x_1 + w_2x_2 +$$
$$w_3x_1^2 + w_4x_1^2x_2 +$$
$$w_5x_1^2x_2^2 + w_6x_1^3x_2 + \ldots)$$

$$C(w) = -\left[\frac{1}{m}\sum_{i=1}^{m}(y^{(i)}log(h_w(x^{(i)})) + (1 - y^{(i)})log(1 - h_w(x^{(i)})))\right]$$
$$+ \frac{\lambda}{2m}\sum_{i=1}^{m}\boxed{w_i^2} \longleftarrow$$

with L2 regularization

Thank you for your attention!