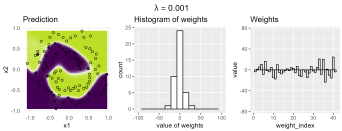
Introduction to Machine Learning

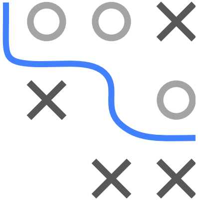
Regularization
Non-Linear Models and Structural Risk Minimization



Learning goals

- Regularization even more important in non-linear models
- Norm penalties applied similarly
- Structural risk minimization

SUMMARY: REGULARIZED RISK MINIMIZATION



If we define (supervised) ML in one line, this might be it:

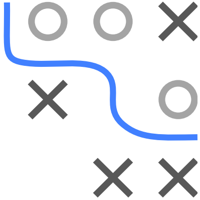
$$\min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \left(\sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right) + \lambda \cdot J(\boldsymbol{\theta}) \right)$$

Can choose for task at hand:

- **hypothesis space** of f, controls how features influence prediction
- **loss** function L, measures how errors are treated
- **regularizer** $J(\theta)$, encodes inductive bias

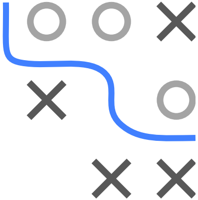
By varying these choices one can construct a huge number of different ML models. Many ML models follow this construction principle or can be interpreted through the lens of RRM.

REGULARIZATION IN NONLINEAR MODELS

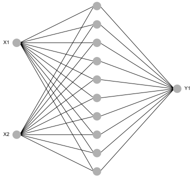


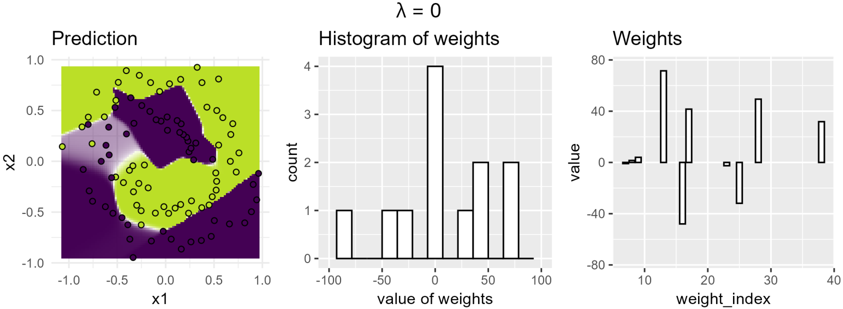
- So far we have mainly considered regularization in LMs
- Can in general also be applied to to non-linear models; vector-norm penalties require numeric params
- Here, we typically use *L*2 regularization, which still results in parameter shrinkage and weight decay
- For non-linear models, regularization is even more important / basically required to prevent overfitting
- Commonplace in methods such as NNs, SVMs, or boosting
- Prediction surfaces / decision boundaries become smoother

REGULARIZATION IN NONLINEAR MODELS



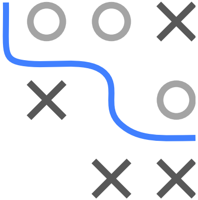
Classification for `spirals` data.
NN with single hidden layer, size 10, $L2$ penalty:



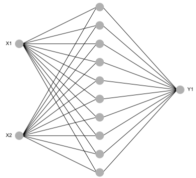


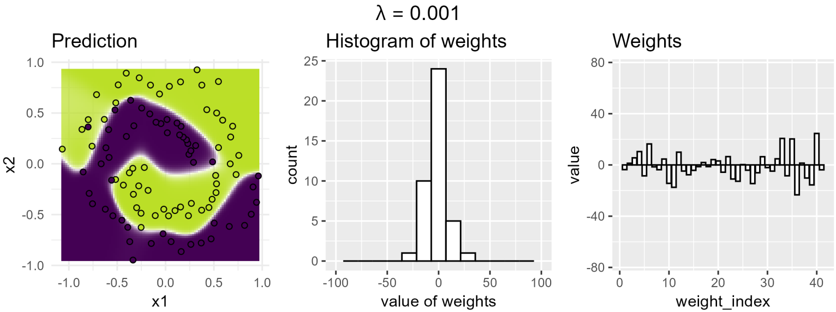
λ affects smoothness of decision boundary and magnitude of weights

REGULARIZATION IN NONLINEAR MODELS



Classification for `spirals` data.
NN with single hidden layer, size 10, *L*2 penalty:



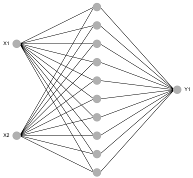


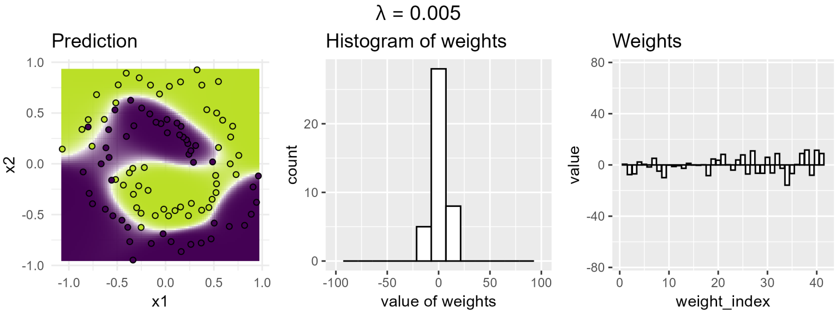
λ affects smoothness of decision boundary and magnitude of weights

REGULARIZATION IN NONLINEAR MODELS

Classification for `spirals` data.

NN with single hidden layer, size 10, *L*2 penalty:

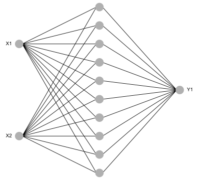


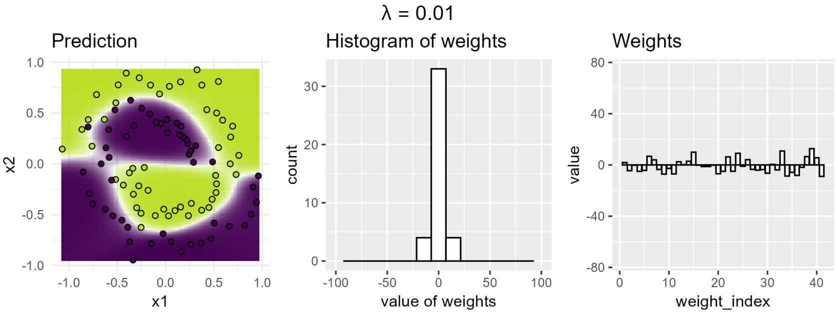


λ affects smoothness of decision boundary and magnitude of weights

REGULARIZATION IN NONLINEAR MODELS

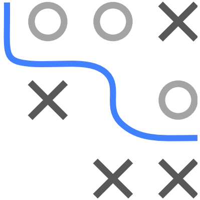
Classification for `spirals` data.
NN with single hidden layer, size 10, *L*2 penalty:





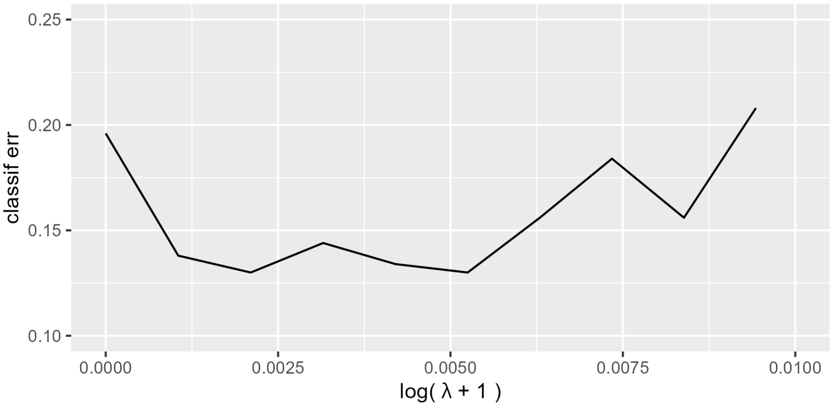
λ affects smoothness of decision boundary and magnitude of weights

REGULARIZATION IN NONLINEAR MODELS



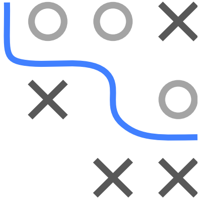
Prevention of overfitting can also be seen in CV.

Same settings as before, but each λ is evaluated with 5x10 REP-CV

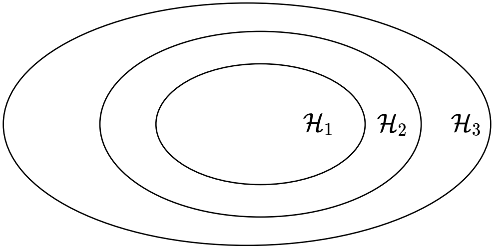


Typical U-shape with sweet spot between overfitting and underfitting

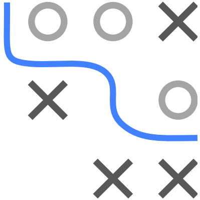
STRUCTURAL RISK MINIMIZATION



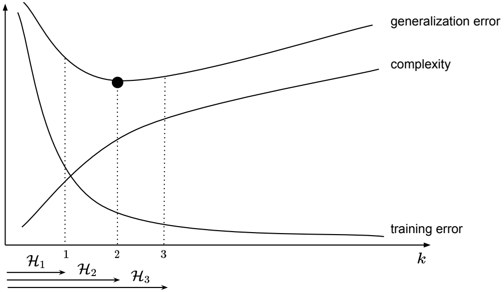
- Can also see this as an iterative process;
  more a “discrete” view on things
- SRM assumes that $\mathcal{H}$ can be decomposed into increasingly complex hypotheses: $\mathcal{H} = \cup_{k \geq 1} \mathcal{H}_k$
- Complexity parameters can be, e.g. the degree of polynomials in linear models or the size of hidden layers in neural networks



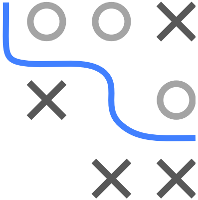
STRUCTURAL RISK MINIMIZATION / 2



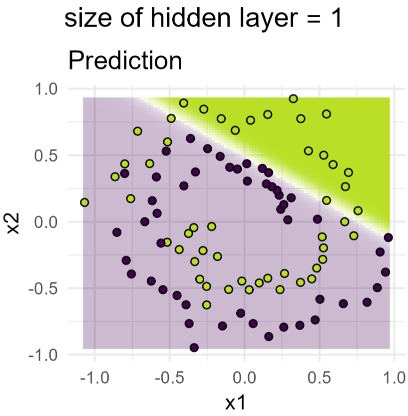
- SRM chooses the smallest k such that the optimal model from $\mathcal{H}_k$ found by ERM or RRM cannot significantly be outperformed by a model from a $\mathcal{H}_m$ with $m > k$
- Principle of Occam's razor
- One challenge might be choosing an adequate complexity measure, as for some models, multiple exist

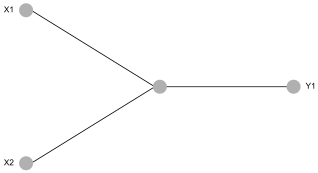


STRUCTURAL RISK MINIMIZATION



Again `spirals`.
NN with 1 hidden layer, and fixed (small) L2 penalty.



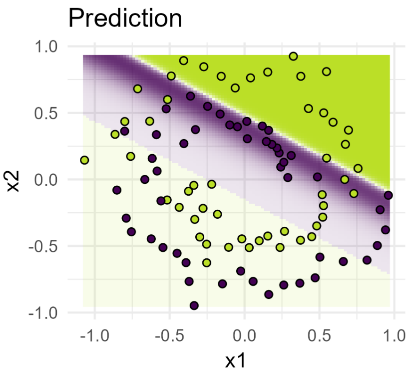


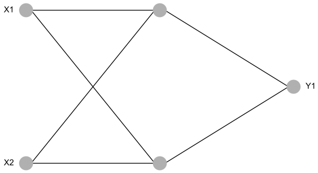
Size affects complexity and smoothness of decision boundary

STRUCTURAL RISK MINIMIZATION

Again `spirals`.
NN with 1 hidden layer, and fixed (small) L2 penalty.

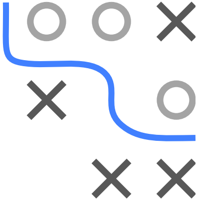
size of hidden layer = 2



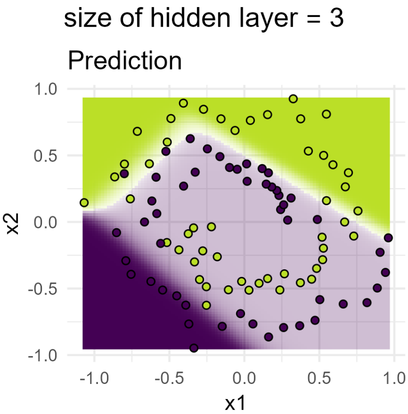


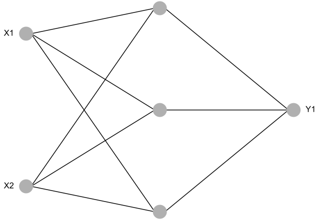
Size affects complexity and smoothness of decision boundary

STRUCTURAL RISK MINIMIZATION



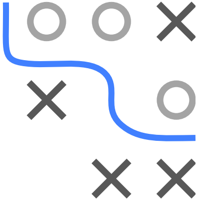
Again `spirals`.
NN with 1 hidden layer, and fixed (small) L2 penalty.



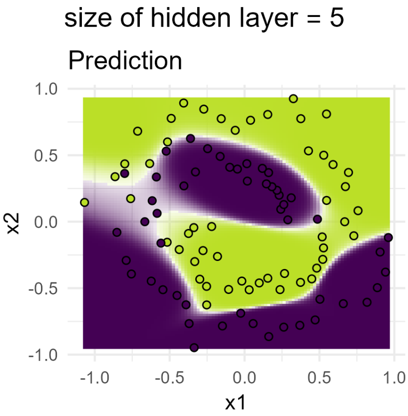


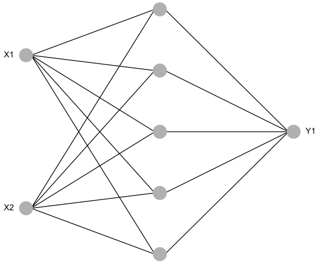
Size affects complexity and smoothness of decision boundary

STRUCTURAL RISK MINIMIZATION



Again `spirals`.
NN with 1 hidden layer, and fixed (small) L2 penalty.



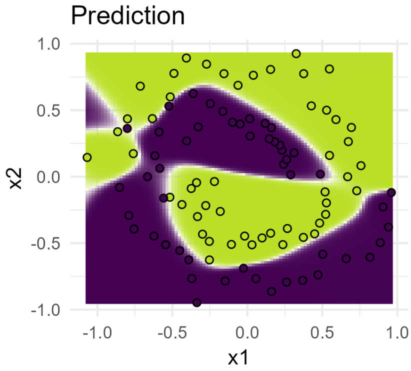


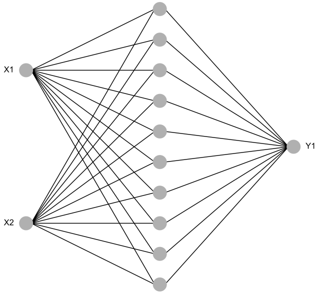
Size affects complexity and smoothness of decision boundary

STRUCTURAL RISK MINIMIZATION

Again `spirals`.
NN with 1 hidden layer, and fixed (small) L2 penalty.

size of hidden layer = 10

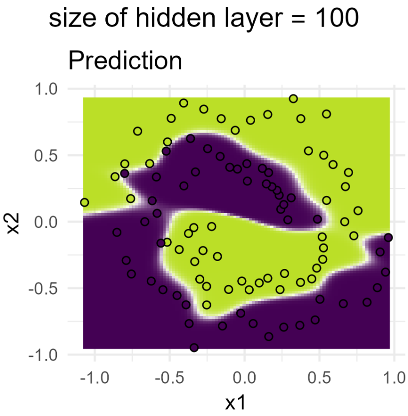


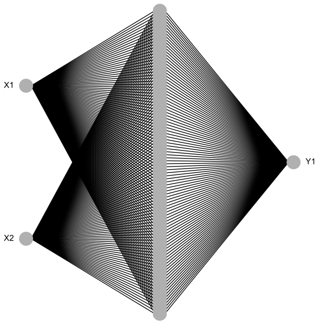


Size affects complexity and smoothness of decision boundary

STRUCTURAL RISK MINIMIZATION

Again `spirals`.
NN with 1 hidden layer, and fixed (small) L2 penalty.

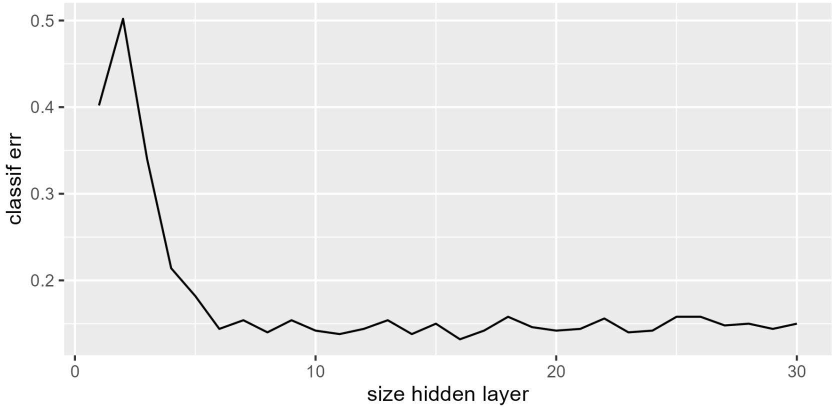




Size affects complexity and smoothness of decision boundary

STRUCTURAL RISK MINIMIZATION

Again, complexity vs CV score.



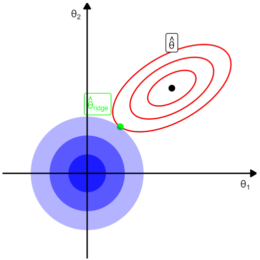
Minimal model with good generalization seems to size=10

STRUCTURAL RISK MINIMIZATION AND RRM

RRM can also be interpreted through SRM,
if we rewrite it in constrained form:

$$\min_{\boldsymbol{\theta}} \quad \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)$$

$$\text{s.t.} \quad \|\boldsymbol{\theta}\|_2^2 \leq t$$



Can interpret going through λ from large to small as through t from small to large. Constructs series of ERM problems with hypothesis spaces $\mathcal{H}_\lambda$, where we constrain norm of $\boldsymbol{\theta}$ to unit balls of growing sizes.