

IAML: Generalization

Chris Williams and Victor Lavrenko
School of Informatics

Semester 1

- ▶ Generalization error
- ▶ Estimating generalization error
- ▶ Example: polynomial regression
- ▶ Under- and over-fitting
- ▶ Cross-validation
- ▶ Regularization
- ▶ Reading: W & F §5.1, 5.3,

1 / 14

2 / 14

Generalization error

- ▶ The real aim of supervised learning is to do well on test data that is not known during training

$$E_{train} = \frac{1}{n} \sum_{i=1}^n error(f_D(\mathbf{x}_i), y_i)$$

$$E_{gen} = \int error(f_D(\mathbf{x}), y(\mathbf{x})) p(\mathbf{x}) d\mathbf{x}$$

where $p(\mathbf{x})$ is the probability density of the input data and $f_D(\mathbf{x})$ is the predictor obtained after training on dataset D .

- ▶ Generalization not memorization
- ▶ E_{gen} is a theoretical quantity
- ▶ Often $E_{gen} > E_{train}$, because the model has been fitted using the training data

Estimating the generalization error

- ▶ Labelled data \rightarrow training data + validation data
- ▶ Train on the training set $\rightarrow f_D$
- ▶ *Estimate* generalization error using the validation set

$$E_{val} = \frac{1}{V} \sum_{v=1}^V error(f_D(\mathbf{x}_v), y_v)$$

sum runs over the V validation patterns

- ▶ E_{val} is an *unbiased*¹ estimator of E_{gen}

¹Here unbiased is used in a statistical sense, i.e. that the expected value of E_{val} is E_{gen} . Do not confuse this with the notion of *inductive bias*.

3 / 14

4 / 14

$$\phi(x) = (1, x, x^2, \dots, x^M)^T$$

- ▶ Choosing values of the parameters that minimize the training error may not lead to the best generalization performance
- ▶ We want the learning machine to model the true regularities in the data, and to ignore noise
- ▶ It is intuitively obvious that you can only expect a model to generalize well if it explains the data surprisingly well given the complexity of the model
- ▶ If the model has as many degrees of freedom as the data it can fit it perfectly, but so what?
- ▶ There is a lot of theory about how to measure model complexity and how to control it to optimize generalization

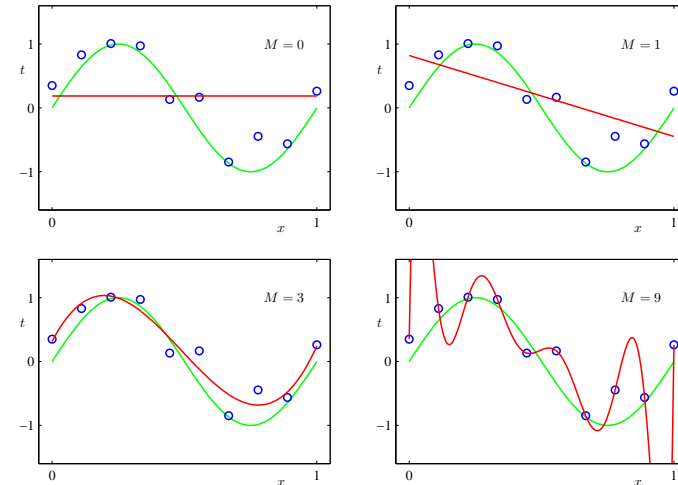
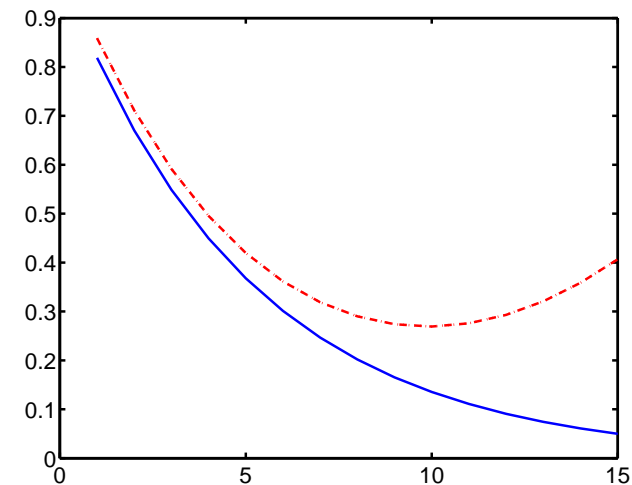


Figure credit: Chris Bishop, PRML

- ▶ Model too simple: underfitting
- ▶ Model too complex: overfitting.
- ▶ Overfitting: A hypothesis $f \in \mathcal{F}$ is said to **overfit** the data if there exists some alternative hypothesis $f' \in \mathcal{F}$ such that f has a smaller training error than f' , but f' has a smaller generalization error than f .
- ▶ Need to balance between under and overfitting: use validation set, or better, cross-validation



Training error (blue) and generalization error (red) with increasing model power

- ▶ Labelled data → training data + validation data
 - ▶ Test set: to check how the final model performs
1. Decide on a set of models to test (e.g. a set of polynomial model orders)
 2. Learn the parameters for all these models (e.g. using maximum likelihood learning)
 3. Check the performance of each model with the learned parameters on the validation set.
 4. Pick the model which performs best on the validation set
 5. Test it on the test set to see how well you should expect it to perform

9 / 14

Regularization

- ▶ If we penalize polynomials that have large values for their coefficients we will get less wiggly solutions

$$\tilde{E}(\mathbf{w}) = |\mathbf{y} - \Phi\mathbf{w}|^2 + \lambda|\mathbf{w}|^2$$

- ▶ Solution is

$$\hat{\mathbf{w}} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T\mathbf{y}$$

- ▶ This is known as *ridge regression*
- ▶ Rather than using a discrete control parameter like M (model order) we can use a continuous parameter λ

11 / 14

- ▶ The idea of holding out a separate validation set seems rather wasteful of data → k -fold cross validation.
- ▶ Divide the labelled data into k parts (or folds), train on $k - 1$ folds, and validate on one. Do this k times, holding out a different fold each time. Common choices for k are 3 or 10
- ▶ Validation performance is average of validation performance on each of the k folds
- ▶ If $k = n$, then we have leave-one-out cross validation (LOO-CV)

10 / 14

The effect of regularization for $M = 9$

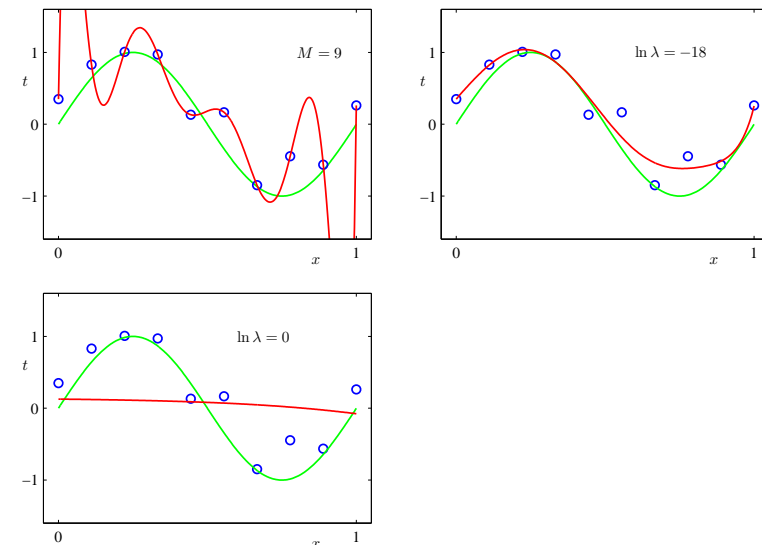
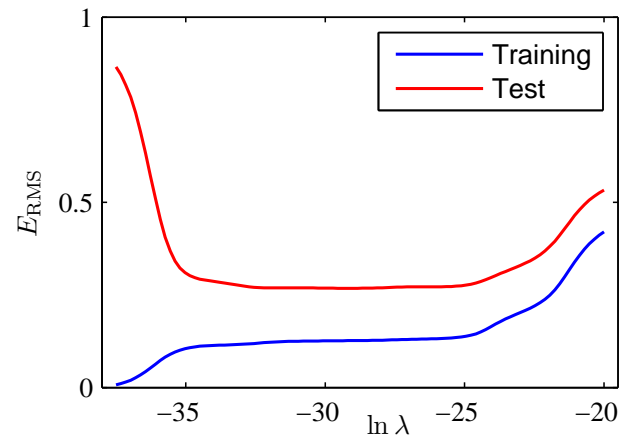


Figure credit: Chris Bishop, PRML

12 / 14

Summary

$M = 9$



Chris Bishop, PRML

- ▶ Generalization error vs training error
- ▶ Under- and over-fitting
- ▶ Estimate generalization error with a validation set (or CV)
- ▶ Regularization
- ▶ There is a lot more theory on the issue of generalization, e.g. Bayesian methods (MLPR)