



UPPSALA
UNIVERSITET

Machine learning – Block 1(b)

Måns Magnusson
Department of Statistics, Uppsala University

Autumn 2022

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation



UPPSALA
UNIVERSITET

This weeks lectures

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

- Regularization
- Model Selection and Assesment
- Cross-Validation
- Evaluate classification models



UPPSALA
UNIVERSITET

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Section 1

Predictive Performance



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

- In the past, tools for assessing models, e.g.:
 - Residuals
 - Leverage, Cook's distance
 - p-values
 - R^2
 - AIC
- Model diagnoses and how well the model fits the data.
- In statistics: focus on estimation and attribution.



Predictive Performance

- Supervised learning: focus on **predictive performance**:
- How well our model \hat{f} trained on

$$\mathcal{T} = \{(y_i, x_i), i = 1, \dots, n\}$$

work when predicting a new observation y_0 from the data generating process $P_{y,x}$.

$$\mathbb{E}(L(Y_0, X_0)|\mathcal{T}) = \int L(Y_0, \hat{f}(X_0))P_{y,x}d(Y_0, X_0)|\mathcal{T}$$

- ability to perform well on previously unobserved inputs is called **generalization**
- Models can be overly **optimistic**¹:
 - explain training data well
 - poor generalizability
- a phenomenon known as **overfitting**.

¹See e.g. Picard, R.R., Cook, R.D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, **79(387)**, 575–583.



UPPSALA
UNIVERSITET

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Section 2

Measuring Performance



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Loss Functions

- To assess the performance we use the loss function for a new unseen observation y_0 and the prediction of that observation $\hat{f}(x_0)$

$$L(y_0, \hat{f}(x_0))$$

- This is quite general and we choose based L based on what we want the model performe well on.
- Examples:

- Regression problems (squared loss/error):

$$L(y_0, \hat{f}(x_0)) = (y_0 - \hat{f}(x_0))^2$$

- Classification (0-1 loss)

$$L(y_0, \hat{f}(x_0)) = I(y_0 \neq \hat{f}(x_0))$$



- When we predict probabilities $\hat{f}(x_0) = \hat{p}$:

$$L(y_0, \hat{p}) = -(y_0 \log \hat{p}) + ((1 - y_0) \log (1 - \hat{p}))$$

Question: Do you recognize the (cross-entropy) loss function?

- Maximizing the likelihood is the same as minimizing the cross-entropy.
- Multi class cross-entropy over M classes

$$L(\mathbf{y}_0, \hat{\mathbf{p}}) = - \sum_{j=1}^M y_{0,j} \log \hat{p}_j$$

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation



The Confusion Matrix

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

- A common way to present performance in classification is the confusion matrix:

Actual	Prediction	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative ((TN)



The Confusion Matrix: Multi-class

- Predictive Performance
- **Measuring Performance**
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Actual	Prediction			
	a	b	c	
a	T_a	F_{ab}	F_{ac}	and
b	F_{ba}	T_b	F_{bc}	
c	F_{ca}	F_{cb}	T_c	

Actual	Prediction		
	TP	FP	FN
a	T_a	$F_{ba} + F_{ca}$	$F_{ab} + F_{ac}$
...



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+FN+TN)}$$

or

$$\text{Accuracy} = \frac{T_a + T_b + T_c}{N}$$

Question: What is the problem with Accuracy?



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Precision

Of all the predicted positives, how many are **actually positive**?

$$\text{Precision} = \frac{(\text{TP})}{(\text{TP} + \text{FP})}$$

or

$$\text{Precision}_a = \frac{(T_a)}{T_a + F_{ba} + F_{ca}}$$

All **predicted** a : $T_a + F_{ba} + F_{ca}$

If we want one precision estimate for all classes:

1. Macro-average ($\text{Precision}_a, \dots, \text{Precision}_c$)
2. Micro-average (use Table 2)



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Recall

Of all **positives**, how many are predicted correctly (recalled)?

$$\text{Recall} = \frac{(\text{TP})}{(\text{TP} + \text{FN})}$$

and

$$\text{Recall}_a = \frac{(T_a)}{T_a + F_{ab} + F_{ac}}$$

All **true/actual** a : $T_a + F_{ab} + F_{ac}$

If we want one precision estimate for all classes:

1. Macro-average ($\text{Recall}_a, \dots, \text{Recall}_c$)
2. Micro-average (use Table 2)



Sensitivity and specificity

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

and

$$\text{Sensitivity} = \text{Recall of positive class} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \text{Recall of negative class} = \frac{TN}{TN+FP}$$



- Predictive Performance
- **Measuring Performance**
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

F1-score

Harmonic mean of Precision and Recall.

$$F_1 = \frac{2}{\text{Precision}^{-1} + \text{Recall}^{-1}} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Very common performance measurement in practice.

If we want one precision estimate for all classes:

1. Macro-average (F_{1a}, \dots, F_{1c})
2. Micro-average (use Table 2)



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Example

Say that we want to classify spam vs. ham.

	$\hat{f}(x) = 1$	$\hat{f}(x) = 0$
$y = 1$	515	91
$y = 0$	85	569

The cell counts yield us estimates of

1. Accuracy: $\frac{515+569}{515+91+85+569} \approx 0.86$
2. Precision: $\frac{515}{515+85} \approx 0.86$
3. Recall: $\frac{515}{515+91} \approx 0.85$
4. F_1 : $\frac{2 \cdot 0.85 \cdot 0.86}{0.85 + 0.86} \approx 0.855$

In this example, we let $\hat{y}_i = 1$ whenever $\hat{\pi}_i > 0.5$.

What if we choose another cut-off level $\hat{\pi}_i > \alpha$ instead?



- Predictive Performance
- **Measuring Performance**
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Classification tables

$\alpha = 0.5$	$\hat{f}(x) = 1$	$\hat{f}(x) = 0$
$y = 0$	515	91
$y = 1$	85	569

Now let $\alpha = 0.3$ instead, so that we are more prone to say that $\hat{y} = 1$:

$\alpha = 0.3$	$\hat{f}(x) = 1$	$\hat{f}(x) = 0$
$y = 1$	462	144
$y = 0$	38	616

The cell counts yield us estimates of

1. Accuracy: $\frac{462+616}{462+38+144+616} \approx 0.86$
2. Precision: $\frac{462}{462+38} \approx 0.92$
3. Recall: $\frac{462}{462+144} \approx 0.76$
4. F_1 : $\frac{2 \cdot 0.92 \cdot 0.76}{0.92 + 0.76} \approx 0.83$

The Precision has increased, but the Recall has decreased...



A more problematic example

A highly unbalanced example. 1001 ham and 17 spam.

Our new classifier: Everything is ham!

	$\hat{f}(x) = 1$	$\hat{f}(x) = 0$
$y = 1$	1001	0
$y = 0$	17	0

The cell counts yield us estimates of

1. Accuracy: $\frac{1001}{1001+17} \approx 0.99$
2. Precision: $\frac{1001}{1001+0} \approx 1.0$
3. Recall: $\frac{1001}{1001+17} \approx 0.99$
4. F_1 : $\frac{2 \cdot 1 \cdot 0.99}{0.99+1} \approx 0.99$
5. But Specificity is 0!



UPPSALA
UNIVERSITET

- Predictive Performance
- Measuring Performance
- **Test and training error**
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Section 3

Test and training error



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

- The main error of interest - *generalization error*
- Conditional Test Error
(Model performance for the **actual** training data):

$$\text{Err}_{\mathcal{T}} = \mathbb{E}_{Y,X}(L(Y_0, \hat{f}(X_0)|\mathcal{T}))$$

- Expected Test Error
(Model performance over **different** training datasets):

$$\text{Err} = \mathbb{E}_{\mathcal{T}}(\mathbb{E}_{Y,X}(L(Y_0, \hat{f}(X_0))))$$

- Sometimes referred to as **generalization** error.



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

- The loss function the algorithm try to **minimize**
- The Error in the training data:

$$\overline{\text{err}} = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i))$$

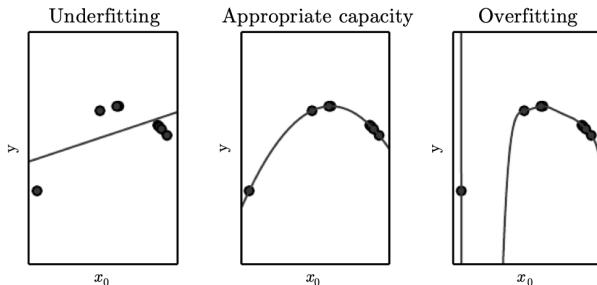


- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Model complexity/capacity

- Model complexity/capacity: The flexibility of the model.
- Underfitting: Too low capacity of model
- Overfitting: Too high capacity of model
- Example: Polynomial regression with higher order terms

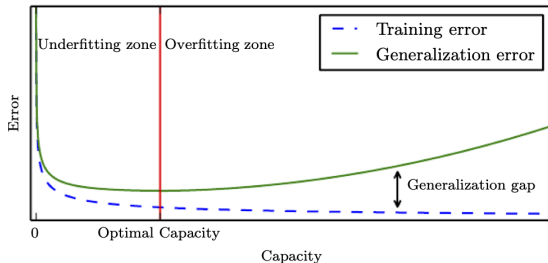
Figure: Model complexity (Goodfellow et al, 2017, Figure 5.2)





Training, test, and complexity

Figure: Test, training, and model complexity (Goodfellow et al, 2017, Figure 5.3)



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation



How to estimate the Test Error (Model Assessment)

- Predictive Performance
- Measuring Performance
- Test and training error
- **Estimating the test error**
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

- We set aside a *test set* from the data
- Use as the last step to *estimate* the test error
- Should only be used **ONCE**
- Size of testset:
 - Common suggestion 10%
 - A statistical estimation problem (choice of sampling size)



Multiple Use of Test Set for Model Assessment

- Predictive Performance
 - Measuring Performance
 - Test and training error
 - Estimating the test error
 - Model Selection
 - Bias and Variance
 - Optimism of Training Error
 - Cross-validation
 - Regularisation
- What happens if we use the test set to pick the model?



UPPSALA
UNIVERSITET

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- **Model Selection**
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Section 5

Model Selection



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- **Model Selection**
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

- We want to select a model based on performance.
- Important when using hyperparameters (as in regularization)



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- **Bias and Variance**
 - Optimism of Training Error
- Cross-validation
- Regularisation

Bias and Variance

Assume we have the following data generating process:

$$Y = f(X) + \epsilon,$$

where $\mathbb{E}(\epsilon) = 0$ and $V(\epsilon) = \sigma_\epsilon$.

$$\begin{aligned}\text{Err}(x_0) &= \mathbb{E}\{(Y - \hat{f}(x_0))^2 | X = x_0\} \\ &= \sigma_\epsilon^2 + \{\mathbb{E}(\hat{f}(x_0)) - f(x_0)\}^2 + \mathbb{E}\{\hat{f}(x_0) - \mathbb{E}(\hat{f}(x_0))\}^2 \\ &= \sigma_\epsilon^2 + \text{Bias}^2(\hat{f}(x_0)) + V(\hat{f}(x_0))\end{aligned}$$

- *Bias*: How close can \hat{f} get to the true model f
- *Variance*: The variability of the predictions from \hat{f}
- *Irreducible/Bayes error*: The best (theoretically) possible model



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- **Bias and Variance**
 - Optimism of Training Error
- Cross-validation
- Regularisation

Bias and Variance: Linear regression

In linear regression we have:

$$\hat{f}(x_i) = \hat{\beta}x_i$$

This give us the following error decomposition:

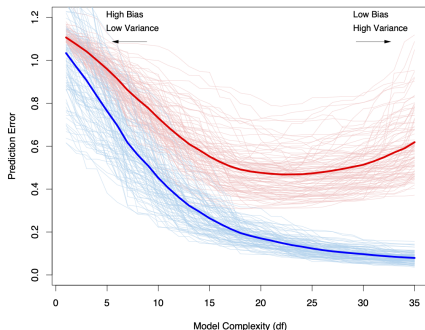
$$\frac{1}{N} \sum_i^N \text{Err}(x_i) = \sigma_\epsilon^2 + \frac{1}{N} \sum_i^N (f(x_i) - E(\hat{f}(x_i)))^2 + \frac{p}{N} \sigma_\epsilon^2$$



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- **Bias and Variance**
 - Optimism of Training Error
- Cross-validation
- Regularisation

Bias and Variance

Figure: Test, training, and model complexity (Hastie et al, 2009, Figure 7)



- High Bias: Underfit
- High Variance: Overfit
- High Irreducible error: No model is good



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Optimism of Training Error

The in-sample test error:

$$\text{Err}_{\text{in}} = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{Y^0} \{ L(Y_i^0, \hat{f}(x_i)) | \mathcal{T} \},$$

where $Y_{0,i}$ is a new response variable condition on x_i .

We have that

$$\mathbb{E}_{\mathbf{y}}(\text{Err}_{\text{in}}) = \mathbb{E}_{\mathbf{y}}(\overline{\text{err}}) + \underbrace{\frac{2}{N} \sum_{i=1}^N \text{Cov}(\hat{y}_i, y_i)}_{\text{optimism}},$$

where $\overline{\text{err}}$ is the training error.

How could we create an optimistic classifier for the training data?



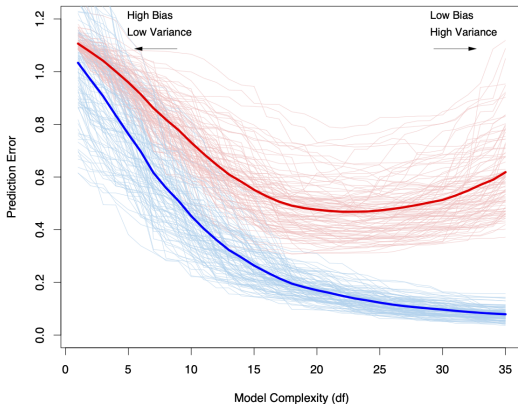
- Predictive Performance
 - Measuring Performance
 - Test and training error
 - Estimating the test error
 - Model Selection
 - Bias and Variance
 - Optimism of Training Error
 - Cross-validation
 - Regularisation
- Under certain conditions we can estimate this optimism.
 - AIC, BIC etc are examples of this – asymptotic predictive performance.



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Find the Optimism!

Figure: Test, training, and model complexity (Hastie et al, 2009, Figure 7)

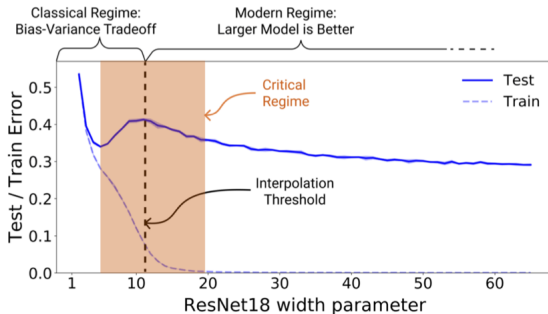




- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

The double descent behaviour of large models

Figure: The double descent of large models (Nakkiran et al., 2019, Figure 1)





UPPSALA
UNIVERSITET

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- **Cross-validation**
- Regularisation

Section 7

Cross-validation



We want to estimate Err for different models and to choose the best model.

Cross-Validation is probably the most popular approach to estimate Err.

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- **Cross-validation**
- Regularisation

- The model is judged only on how well it does predictions for new data.
- No need for rules-of-thumbs to verify that tests and estimators are applicable.
- No need to worry about significance levels, standard errors etc.
- Equally useful for frequentist, Bayesian and algorithmic methods (and these can easily be compared).



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- **Cross-validation**
- Regularisation

Cross-Validation Algorithm

Figure: Cross-Validation (Hastie et al, 2009, p. 222, 242)



1. Split data in K folds
2. For each fold $k = 1, 2, \dots, K$
 - 2.1 Use all samples except those in k to build the predictive model
 - 2.2 Use the model and predict the observations in fold k

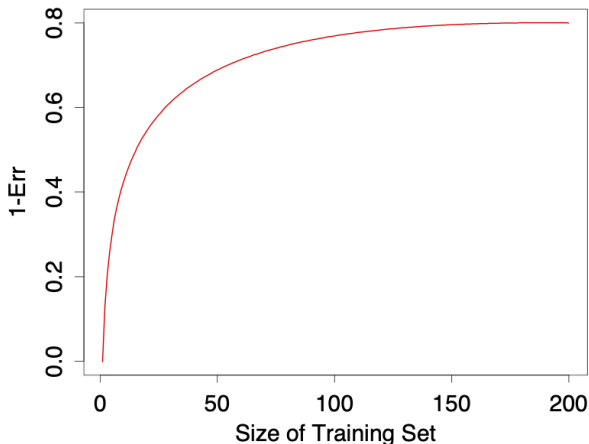
$$CV(\hat{f}, \alpha) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{y}_{-\kappa(i)}(x_i, \alpha))$$



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- **Cross-validation**
- Regularisation

The Bias of Cross-Validation

Figure: Cross-Validation Bias (Hastie et al, 2009, Fig. 7.8)





- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

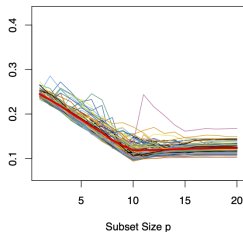
- We set $K = N$
- Benefits
 - Almost unbiased estimate of Err
 - Can be less computationally costly in some situations
- Drawbacks
 - Higher Variance
 - Can be more computationally costly (naive implementation)



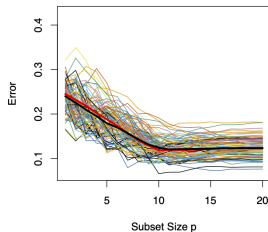
Leave-One-Out Cross Validation

Figure: Cross-Validation Bias (Hastie et al, 2009, Fig. 7.14)

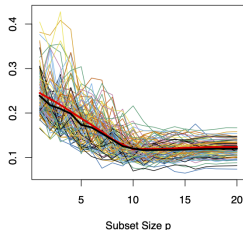
Prediction Error



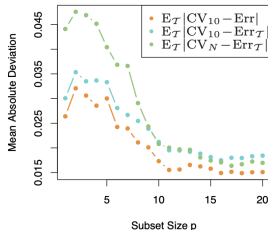
10-Fold CV Error



Leave-One-Out CV Error



Approximation Error



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation



The role of the data generating process

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

- we assume that testset and train set are different observations from the same data generating process

$$\mathbf{d} = \{(y_i, \mathbf{x}_i), i = 1, \dots, n\} \sim P_{y, \mathbf{x}}$$

- Things that can go wrong:
 - temporal leak/concept drift
 - duplicated observations
 - non-randomized data
- Example: Evaluating prediction models for Elections



UPPSALA
UNIVERSITET

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Section 8

Regularisation



- Predictive Performance
 - Measuring Performance
 - Test and training error
 - Estimating the test error
 - Model Selection
 - Bias and Variance
 - Optimism of Training Error
 - Cross-validation
 - Regularisation
- Linear regression and logistic regression are examples of **generalised linear models**, GLMs.
 - Both use maximum likelihood estimation for fitting the model, where the likelihood function $L(\beta)$ is maximised.



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Regularised regression models

- In some situations, for instance when the predictors are highly collinear, when there are too many predictors or when there is complete separation in the data, maximum likelihood estimation is unstable.
 - Either the solution is not unique, or minuscule changes in the data can change the solution completely.
 - Such datasets are increasingly common in e.g. genomics, finance, astronomy and image analysis.
- In such cases, **regularisation/shrinkage methods** can be used instead.
- In a regularized GLM, it is not the likelihood $L(\beta)$ that is maximized, but a **regularised** function $L(\beta) \cdot p(\beta)$, where p is a penalty function that typically forces the resulting estimates to be closer to 0, which leads to a stable solution.



Regularised regression models

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Regularised linear regression models increase the **bias** of the estimates, but lowers their **variance**, thereby potentially decreasing the MSE.



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Connection to Bayesian estimation

In Bayesian estimation, a **prior distribution** $p(\beta)$ for the parameters β_i is chosen.

The estimates are then computed from the conditional distribution of the β_i given the data, called the **posterior distribution**.

Using Bayes' theorem, we find that

$$P(\beta|\mathbf{x}) \propto L(\beta) \cdot p(\beta),$$

i.e. that the posterior distribution is proportional to the likelihood times the prior.

A special type of Bayesian estimator is the **maximum a posteriori (MAP)** estimator, which is found by maximizing the above expression (i.e. finding the mode of the posterior).

This is equivalent to the estimates from a regularised frequentist model with penalty function $p(\beta)$!



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Inference and invariance

- Regularised regression models are not invariant under linear rescaling of the predictors.
 - If a predictor is multiplied by a scalar $a \neq 0$, this can change the entire model.
 - A model with measurements in inches might yield completely different results from a model with measurements in cm.
- For this reason, it is widely agreed that the predictors should be standardized to have mean 0 and variance 1 before a regularised model is fitted.
 - With this approach we choose a particular (natural?) scaling, among all possible scalings.
 - All predictors are on the same scale and are therefore treated equally by the penalty function.



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

- Hypothesis tests are available (e.g. Lockhart et al. (2014), A significance test for the lasso, *Annals of Statistics*), but I advise against using them.
- Note that the hypothesis tests will be conditioned on the choice of scaling.
 - Because of this, regularised models are not appropriate for hypothesis testing – the p-values could change completely if we rescaled the data!
- Regularised regression models are however very useful for predictive modelling.



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

L_q -penalties

The most popular penalty terms correspond to common L_q -norms. On a log-scale, the function to be maximized is

$$\ell(\beta) + \lambda \sum_{i=1}^p |\beta_i|^q,$$

where $\ell(\beta)$ is the loglikelihood of β and $\sum_{i=1}^p |\beta_i|^q$ is the L_q -norm, with $q \geq 0$.

This is equivalent to maximizing $\ell(\beta)$ under the constraint that $\sum_{i=1}^p |\beta_i|^q \leq \frac{1}{h(\lambda)}$, for some increasing positive function h .

- Relies on the **sparsity** assumption that most β are 0.

$\lambda \geq 0$ is a **smoothing parameter**:

- When $\lambda = 0$, we are back at the standard ML-estimate.
- The $\hat{\beta}$ are forced to be closer to 0 when λ increases.
- λ is usually chosen using cross-validation.



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Ridge regression

When the L_2 penalty is used, the regularised model is called **ridge regression**, for which we maximize

$$\ell(\beta) + \lambda \sum_{i=1}^p \beta_i^2.$$

- Invented and reinvented by several authors, from the 1940's onwards.
- In a linear model, the OLS estimate is $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, whereas the ridge estimate is $\hat{\beta} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$. The $\lambda \mathbf{I}$ is the 'ridge'.
- The β_i can become very small, but are never pushed all the way down to 0.
- In a Bayesian context, this corresponds to putting a standard normal prior on the β_i .



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Lasso

When the L_1 penalty is used, the regularised model is called the **lasso** (Least Absolute Shrinkage and Selection Operator), for which we maximize

$$\ell(\beta) + \lambda \sum_{i=1}^p |\beta_i|.$$

- Introduced by Robert Tibshirani in 1996.
- As λ increases, more and more β_i become 0.
 - Simultaneously performs estimation and variable selection!
- In a Bayesian context, this corresponds to putting a standard Laplace prior on the β_i .



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Functions for regularised generalized linear models (linear, logistic, Poisson, multinomial, and more) are available e.g. in the `glmnet` package for R.

The syntax used is somewhat different from that for `glm` and `lm`.



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Model Selection
- Bias and Variance
 - Optimism of Training Error
- Cross-validation
- Regularisation

Generalizations

Regularised models have been a hot research topics in the last 20 years. Some additional important models are:

- **Elastic net:** a compromise between ridge and lasso, in which

$$\ell(\beta) + \lambda_1 \sum_{i=1}^p |\beta_i| + \lambda_2 \sum_{i=1}^p \beta_i^2$$

is maximized.

- Introduced by Zou and Hastie in 2005.
 - Is better than the lasso at handling correlated predictors.
 - Has two smoothing parameters that we need to choose.
 - Available in the `glmnet` package.
- **Group lasso:** a version of the lasso in which variables can be grouped before fitting the model. The group lasso then selects groups of variables rather than individual variables.
 - Introduced by Yuan and Lin in 2006.
 - Useful e.g. when we have dummies for categorical variables (in contrast, the lasso may choose to only include the dummies for some of the categories).