



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
- Bias and Variance
- Cross-validation



UPPSALA
UNIVERSITET

This weeks lectures

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

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



UPPSALA
UNIVERSITET

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

Section 1

Predictive Performance



- Predictive Performance
 - Measuring Performance
 - Test and training error
 - Estimating the test error
 - Bias and Variance
 - Cross-validation
- In the past, tools for assessing models, e.g.:
 - Residuals
 - Leverage, Cook's distance
 - p-values
 - R^2
 - AIC
 - (LOO-CV)
 - Model diagnoses and how well the model fits the data.



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

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



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

How well our model $\hat{f}_{\mathcal{T}}$ trained on

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

work when predicting a **new observation** (y_0, x_0) from the data generating process $P_{y,x}$.

$$\mathbb{E} \left[L(y_0, \hat{f}_{\mathcal{T}}(x_0)) \right] = \int L(y_0, \hat{f}_{\mathcal{T}}(x_0)) P_{(y,x)} d(y_0, x_0)$$

where $L(y, x)$ is a **loss function** (e.g. $L(x, y) = (x - y)^2$)



Predictive Performance

- Predictive Performance
 - Measuring Performance
 - Test and training error
 - Estimating the test error
 - Bias and Variance
 - Cross-validation
- ability to perform well on previously unobserved inputs is called **generalization**
 - $\mathbb{E} \left[L(y_0, \hat{f}_{\mathcal{T}}(x_0)) \right]$ is the **generalization error**



Predictive Performance

- Predictive Performance
 - Measuring Performance
 - Test and training error
 - Estimating the test error
 - Bias and Variance
 - Cross-validation
- ability to perform well on previously unobserved inputs is called **generalization**
 - $\mathbb{E} \left[L(y_0, \hat{f}_{\mathcal{T}}(x_0)) \right]$ is the **generalization error**
 - Models can be **overfit**
 - explain training data well
 - poor generalizability



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

- ability to perform well on previously unobserved inputs is called **generalization**
- $\mathbb{E} \left[L(y_0, \hat{f}_{\mathcal{T}}(x_0)) \right]$ is the **generalization error**
- Models can be **overfit**
 - explain training data well
 - poor generalizability
- Models can be **underfit**
 - explain training data poorly
 - poor generalizability



UPPSALA
UNIVERSITET

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

Section 2

Measuring Performance



Loss Functions

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

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

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation



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.

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation



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 perform well on.
- Examples:
 - Regression problems (squared loss/error):

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

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

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)

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



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

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)

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

- In general: The **negative log likelihood** is a good loss function



Cross-Entropy Loss

- 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?

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation



- 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.

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

Cross-Entropy Loss

- 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

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



The Confusion Matrix

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

- 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
- Bias and Variance
- Cross-validation

		Prediction			
	Actual	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	
		Prediction			
Actual		TP	FP	FN	
a		T_a	$F_{ba} + F_{ca}$	$F_{ab} + F_{ac}$	and N is the
...		
		sum over all cells.			



UPPSALA
UNIVERSITET

Accuracy

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

or

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

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



UPPSALA
UNIVERSITET

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

Accuracy

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

or

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

Question: Can you see a problem with Accuracy?



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

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)



UPPSALA
UNIVERSITET

Recall

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

- Predictive Performance
- **Measuring Performance**
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

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
- Bias and Variance
- Cross-validation

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

and

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



Harmonic mean of Precision and Recall.

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

Question: What happens if Precision or Recalls goes toward zero/one?

- Predictive Performance
- **Measuring Performance**
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation



Harmonic mean of Precision and Recall.

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

Question: What happens if Precision or Recalls goes toward zero/one? 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
- Bias and Variance
- Cross-validation



UPPSALA
UNIVERSITET

Example

Say that we want to classify spam vs. ham.

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

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 : $\frac{2 \cdot 0.85 \cdot 0.86}{0.85 + 0.86} \approx 0.855$



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

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?



UPPSALA
UNIVERSITET

Classification tables

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

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation



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...



UPPSALA
UNIVERSITET

A more problematic example

A highly unbalanced example. 1001 ham and 17 spam.

Our new classifier: Everything is ham!

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation



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

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation



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 : $\frac{2 \cdot 1 \cdot 0.99}{0.99+1} \approx 0.99$



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. F_1 for negative class: $\frac{2 \cdot 0 \cdot 0}{0+0}$ Not defined, but 0 in the limit as Precision and Recall goes to 0.
6. And Specificity is 0!



UPPSALA
UNIVERSITET

Questions?

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

Questions?



UPPSALA
UNIVERSITET

- Predictive Performance
- Measuring Performance
- **Test and training error**
- Estimating the test error
- Bias and Variance
- Cross-validation

Section 3

Test and training error



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

Test Error

- The main error of interest - *generalization error*
- Conditional Test Error
(Performance for the model trained on **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.
- Conditional Test Error is more difficult to estimate than the Expected Test Error (Bates, Hastie, and Tibshiriani, 2021)



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

- Training error: The loss the algorithm try to minimize
- The Error in the training data:

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

where $L(y, x)$ is the loss function.

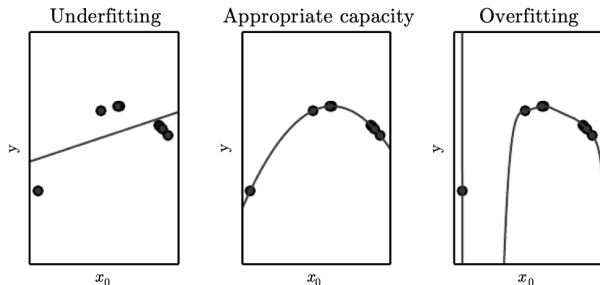


- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

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

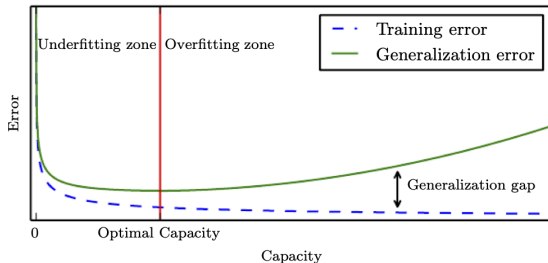
Figure: Model complexity (Goodfellow et al, 2017, Figure 5.2)





Training, test, and complexity

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





How to estimate the Test Error (Model Assessment)

- Predictive Performance
 - Measuring Performance
 - Test and training error
 - Estimating the test error
 - Bias and Variance
 - Cross-validation
- We set aside a **test set** from the data
 - Use as the last step to **estimate** the **generalization error**
 - Should only be used **ONCE** (or a few times)



How to estimate the Test Error (Model Assessment)

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

- We set aside a **test set** from the data
- Use as the last step to **estimate** the **generalization error**
- Should only be used **ONCE** (or a few times)
- Size of testset:
 - Common suggestion is 10%, but
 - It is 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
 - Bias and Variance
 - Cross-validation
- 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**
- Bias and Variance
- Cross-validation

Questions?

Questions?



UPPSALA
UNIVERSITET

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- **Bias and Variance**
- Cross-validation

Section 5

Bias and Variance



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- **Bias and Variance**
- Cross-validation

Bias and Variance

Assume we have the following data generating process:

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

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

We have an estimated model \hat{f} and want to predict a new, unseen, observation x_0 . The error can then be decomposed into:

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



Bias and Variance

Assume we have the following data generating process:

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

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

We have an estimated model \hat{f} and want to predict a new, unseen, observation x_0 . The error can then be decomposed into:

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

- *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 minimum possible error



In linear regression we have:

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

This give us the following error decomposition:

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

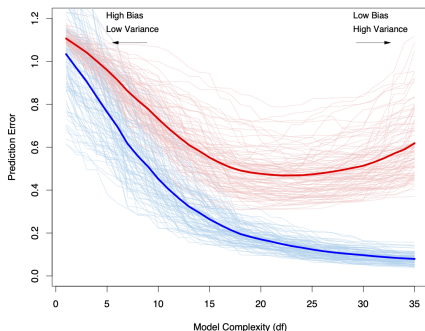
- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- **Bias and Variance**
- Cross-validation



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

Bias and Variance

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

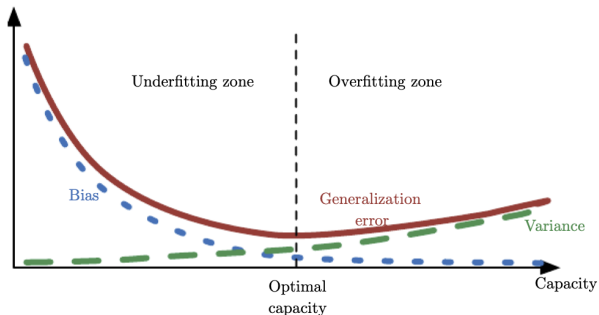


- High Bias: Underfit
- High Variance: Overfit
- High Irreducible error: No model is good



Bias and Variance

Figure: Bias and variance (Goodfellow et al., 2017, Figure 5.6)



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

Optimism

The in-sample **test** error:

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

where Y_i^0 is a **new variable conditioned on x_i** .



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- **Bias and Variance**
- Cross-validation

Optimism

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_i^0 is a **new variable conditioned on x_i** .

We have that

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

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

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



UPPSALA
UNIVERSITET

Estimating Optimism

- Under certain conditions we can estimate this optimism.
- AIC is an example of this – asymptotic predictive performance.

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- **Bias and Variance**
- Cross-validation

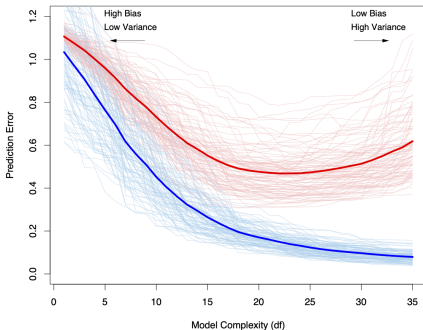


- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

Estimating Optimism

- Under certain conditions we can estimate this optimism.
- AIC is an example of this – asymptotic predictive performance.
- **Question:** What is the optimism in the Figure below?

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

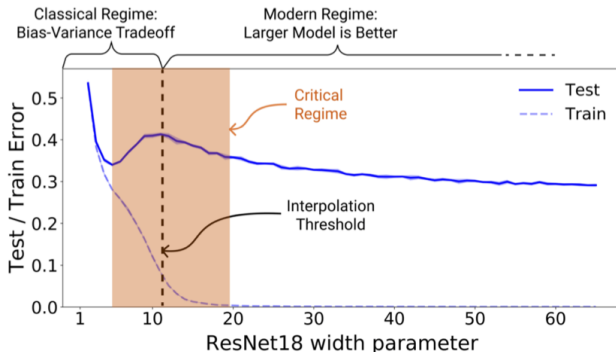




- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

The double descent of large models

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





UPPSALA
UNIVERSITET

Questions?

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- **Bias and Variance**
- Cross-validation

Questions?



UPPSALA
UNIVERSITET

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

Section 6

Cross-validation



Cross-Validation

- We want to estimate **Err** for different models and to **choose the best model** where

$$\begin{aligned}\text{Err} &= \mathbb{E}_{\mathcal{T}}(\text{Err}_{\mathcal{T}}) \\ &= \mathbb{E}_{\mathcal{T}}(\mathbb{E}(L(Y_0, X_0) | \mathcal{T})) \\ &= \int \left(\int L(Y_0, \hat{f}(X_0)) P_{y,x} d(Y_0, X_0) | \mathcal{T} \right) d\mathcal{T}\end{aligned}$$

- Cross-Validation is probably the most popular approach to estimate **Err** and choose between models because it is
 1. Conceptually easy to understand
 2. Easy to implement
 3. No need for rules-of-thumbs to verify that it is applicable
 4. Equally useful for many different type of models
 5. Flexible for the use case at hand



Cross-Validation

- We want to estimate **Err** for different models and to **choose the best model** where

$$\begin{aligned}\text{Err} &= \mathbb{E}_{\mathcal{T}}(\text{Err}_{\mathcal{T}}) \\ &= \mathbb{E}_{\mathcal{T}}(\mathbb{E}(L(Y_0, X_0)|\mathcal{T})) \\ &= \int (\int L(Y_0, \hat{f}(X_0)) P_{y,x} d(Y_0, X_0) | \mathcal{T}) d\mathcal{T}\end{aligned}$$

- Cross-Validation is probably the most popular approach to estimate **Err** and choose between models because it is
 1. Conceptually easy to understand
 2. Easy to implement
 3. No need for rules-of-thumbs to verify that it is applicable
 4. Equally useful for many different type of models
 5. Flexible for the use case at hand
- Common approach to **learn hyper parameters** (that is a model choice)



- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

Cross-Validation Algorithm

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

Train		Validation	Test	
1	2	3	4	5
Train	Train	Validation	Train	Train

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 $\hat{f}(x)$
 - 2.2 Use the model and predict the observations in fold k

$$\text{Err}_{\text{CV}}(\hat{f}) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}_{-\kappa(i)}(x))$$

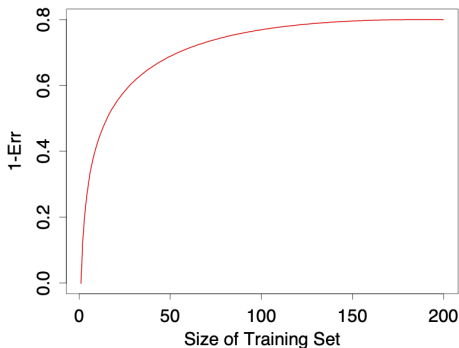


- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

The Bias of Cross-Validation

- Cross-validation estimation of **Err** will be biased
- The training data size is smaller than the full data

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





K-fold Cross Validation

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

- Common K are: $K = \{2, 5, 10\}$
- Smaller K gives larger bias
- Larger K is computationally more costly
- $K = 10$ is a common approach



Leave-One-Out Cross Validation

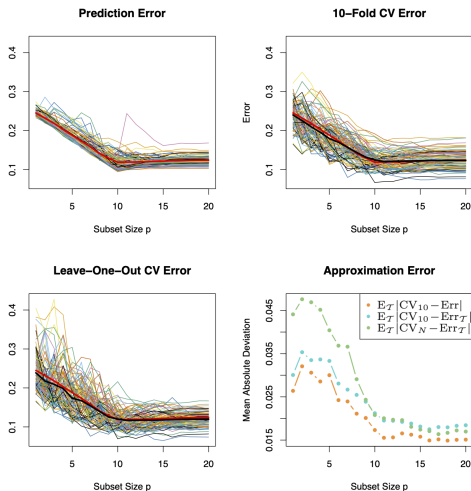
- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

- When $K = N$
- Benefits
 - Almost unbiased estimate of Err
 - Sometimes we only need to train our model once
- Drawbacks
 - Higher Variance in estimate of Err
 - Can be more computationally very costly (naive implementation)
 - Can be unstable/less robust in some settings



Leave-One-Out Cross Validation

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





The role of the data generating process

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

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

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

- The (naive) assumption: independence
- Things that can go wrong:
 - temporal leak/concept drift
 - duplicated observations
 - non-randomized data



The role of the data generating process

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

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

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

- The (naive) assumption: independence
- Things that can go wrong:
 - temporal leak/concept drift
 - duplicated observations
 - non-randomized data

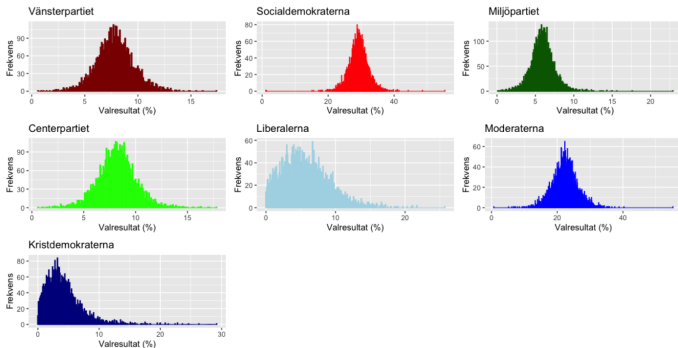


- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

Example: Election prediction

- We want to predict the next election
- We know that there are "concept drift"
- Solution in Frölander and Uddhammar (2021) and Olsson and Ölfvingsson (2021)
 1. LOO-CV on the elections 1973-2014
 2. The elections 2018 as the final validation set

Figure: Predictive distr. (Olsson and Ölfvingsson, 2021, Fig. 6)





UPPSALA
UNIVERSITET

Questions?

- Predictive Performance
- Measuring Performance
- Test and training error
- Estimating the test error
- Bias and Variance
- Cross-validation

Questions?