



UPPSALA
UNIVERSITET

Hur strukturerar man sitt maskininlärningsprojekt?

Möjligheter och fallgorpar

Niklas Wahlström

Avdelningen för systemteknik
Institutionen för informationsteknologi

Docenturföreläsning
Oktober 19, 2023.

MACHINE LEARNING

A First Course for
Engineers and
Scientists

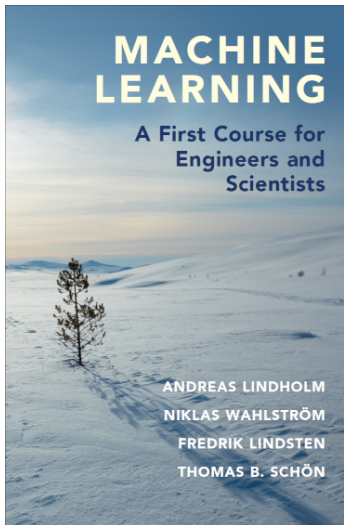
ANDREAS LINDHOLM
NIKLAS WAHLSTRÖM
FREDRIK LINDSTEN
THOMAS B. SCHÖN

Contents

5.6	Hyperparameter Optimisation	295
5.7	Further Reading	299
6	Neural Networks and Deep Learning	
6.1	The Neural Network Model	311
6.2	Training a Neural Network	315
6.3	Convolutional Neural Networks	315
6.4	Dropout	316
6.5	Further Reading	316
6.A	Derivation of the Backpropagation Equations	316
7	Ensemble Methods: Bagging and Boosting	
7.1	Bagging	330
7.2	Random Forests	330
7.3	Boosting and AdaBoost	330
7.4	Gradient Boosting	330
7.5	Further Reading	334
8	Non-linear Input Transformations and Kernels	
8.1	Creating Features by Non-linear Input Transformations	334
8.2	Kernel Ridge Regression	334
8.3	Support Vector Regression	334
8.4	Kernel Theory	334
8.5	Support Vector Classification	334
8.6	Further Reading	334
8.A	The Representer Theorem	334
8.B	Derivation of Support Vector Classification	334
9	The Bayesian Approach and Gaussian Processes	
9.1	The Bayesian Idea	334
9.2	Bayesian Linear Regression	334
9.3	The Gaussian Process	334
9.4	Practical Aspects of the Gaussian Process	334
9.5	Other Bayesian Methods in Machine Learning	334
9.6	Further Reading	334
9.A	The Multivariate Gaussian Distribution	334
10	Generative Models and Learning from Unlabelled Data	
10.1	The Gaussian Mixture Model and Discriminant Analysis	334
10.2	Cluster Analysis	334
10.3	Deep Generative Models	334
10.4	Representation Learning and Dimensionality Reduction	334
10.5	Further Reading	334

Contents

11	User Aspects of Machine Learning	295
11.1	Defining the Machine Learning Problem	295
11.2	Improving a Machine Learning Model	299
11.3	What If We Cannot Collect More Data?	307
11.4	Practical Data Issues	311
11.5	Can I Trust my Machine Learning Model?	315
11.6	Further Reading	316
12	Ethics in Machine Learning	317
12.1	Fairness and Error Functions	317
12.2	Misleading Claims about Performance	322
12.3	Limitations of Training Data	330
12.4	Further Reading	334
	Bibliography	335
	Index	343



Contents

5.6	Hyperparameter Optimisation	295
5.7	Further Reading	295
6	Neural Networks and Deep Learning	
6.1	The Neural Network Model	299
6.2	Training a Neural Network	307
6.3	Convolutional Neural Networks	311
6.4	Dropout	315
6.5	Further Reading	316
6.A	Derivation of the Backpropagation Equations	
7	Ensemble Methods: Bagging and Boosting	
7.1	Bagging	317
7.2	Random Forests	322
7.3	Boosting and AdaBoost	330
7.4	Gradient Boosting	334
7.5	Further Reading	
8	Non-linear Input Transformations and Kernels	
8.1	Creating Features by Non-linear Input Transformations	
8.2	Further Reading	
8.3		
8.4		
8.5		
8.6		
8.A		
8.B		
9	The Multivariate Gaussian Distribution	
9.1		
9.2		
9.3		
9.4		
9.5		
9.6	Further Reading	
9.A	The Multivariate Gaussian Distribution	
10	Generative Models and Learning from Unlabelled Data	
10.1	The Gaussian Mixture Model and Discriminant Analysis	
10.2	Cluster Analysis	
10.3	Deep Generative Models	
10.4	Representation Learning and Dimensionality Reduction	
10.5	Further Reading	

Contents

11	User Aspects of Machine Learning	295
11.1	Defining the Machine Learning Problem	295
11.2	Improving a Machine Learning Model	299
11.3	What If We Cannot Collect More Data?	307
11.4	Practical Data Issues	311
11.5	Can I Trust my Machine Learning Model?	315
11.6	Further Reading	316
12	Ethics in Machine Learning	317
12.1	Fairness and Error Functions	317
12.2	Misleading Claims about Performance	322
12.3	Limitations of Training Data	330
12.4	Further Reading	334
	Bibliography	335
	Index	343

11	User Aspects of Machine Learning	295
11.1	Defining the Machine Learning Problem	295
11.2	Improving a Machine Learning Model	299
11.3	What If We Cannot Collect More Data?	307
11.4	Practical Data Issues	311
11.5	Can I Trust my Machine Learning Model?	315
11.6	Further Reading	316

Vad är maskininlärning?

*Maskininlärning handlar om att konstruera datorprogram som automatiskt kan **lära sig** att lösa komplicerade problem utifrån **data**, utan att vara specifikt programmerade för varje uppgift.*

Ex) God- eller elakartade pigmentförändringar

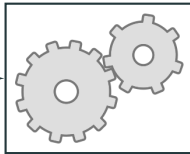
Kan man lära en dator skilja på god- och elakartade leverfläckar?

Bild på leverfläck



Indata

Datorprogram



Modell

godartad eller
elakartad?

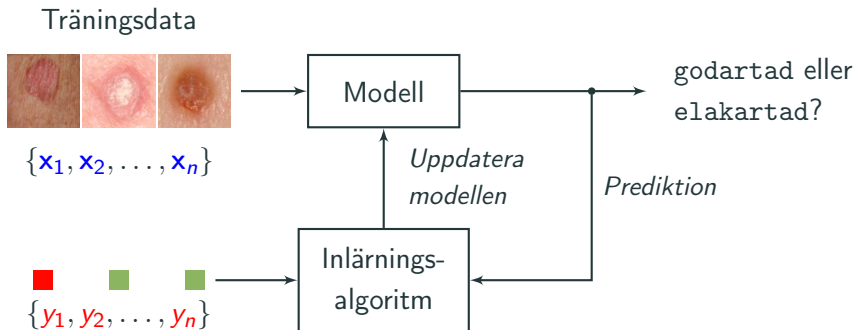
Prediktion

Klassificering: ta fram en **modell** som, för varje datapunkt **x**, kan prediktera dess **klass y** (t.ex. $y \in \{\text{godartad, elakartad}\}$).

Klassificering

Klassificering: ta fram en **modell** som, för varje datapunkt \mathbf{x} , kan prediktera dess **klass** y (t.ex. $y \in \{\text{godartad, elakartad}\}$).

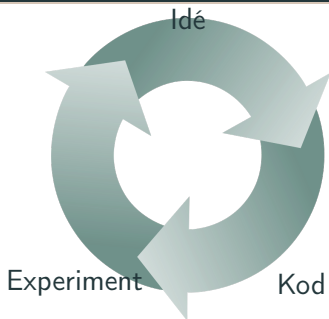
Övervakad inlärning: Modellen **tränas** genom att anpassa den till annoterade **träningsdata** $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$.



Tillämpad maskininlärning är en iterativ process

Finns många val att göra

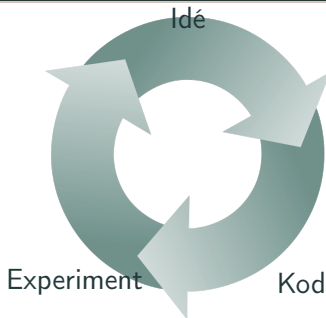
- Vilken modell ska vi välja?
- Hur många parametrar ska modellen ha?
- ...



Tillämpad maskininlärning är en iterativ process

Finns många val att göra

- Vilken modell ska vi välja?
- Hur många parametrar ska modellen ha?
- ...

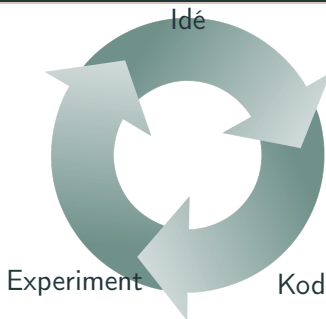


Vad är en bra modell?

Tillämpad maskininlärning är en iterativ process

Finns många val att göra

- Vilken modell ska vi välja?
- Hur många parametrar ska modellen ha?
- ...

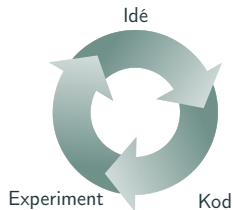


Vad är en bra modell?

Hur kan man förbättra modellen?

Mål: Introducera ett strukturerat sätt att **utvärdera** och **förbättra** en maskininlärningsmodell.

1. Tränings-, validerings- och testdata
 - Fallgrop 1: Dataläckage
2. Utvärderingsmättet
 - Fallgrop 2: Inte ha *ett* utvärderingsmått
3. Träningsfel och generaliseringsfel
 - Fallgrop 3: Att lägga tid på fel saker.



Mål: Introducera ett strukturerat sätt att **utvärdera** och **förbättra** en maskininlärningsmodell.

1. Tränings-, validerings- och testdata

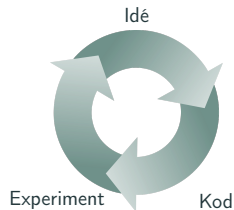
- Fallgrop 1: Dataläckage

2. Utvärderingsmättet

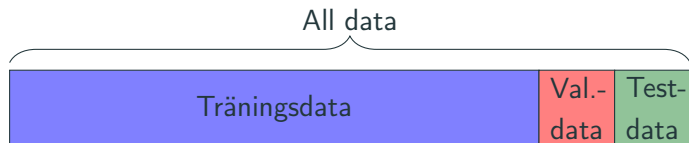
- Fallgrop 2: Inte ha *ett* utvärderingsmått

3. Träningsfel och generaliseringsfel

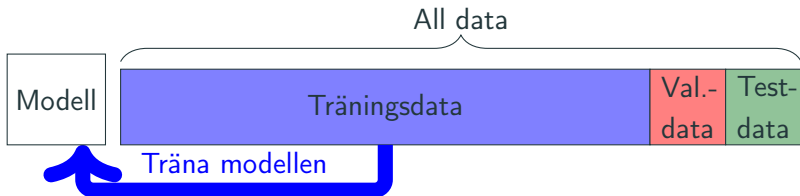
- Fallgrop 3: Att lägga tid på fel saker.



Tränings-, validerings-, och testdata

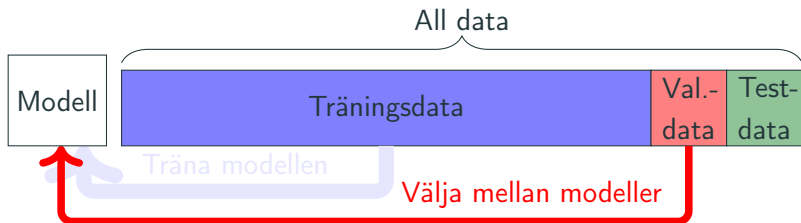


Tränings-, validerings-, och testdata



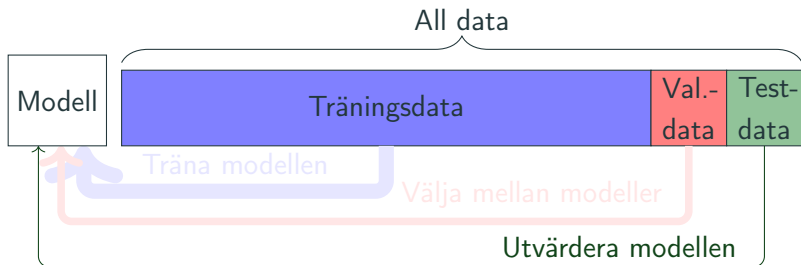
- **Träningsdata:** Används för att **träna** modellen

Tränings-, validerings-, och testdata



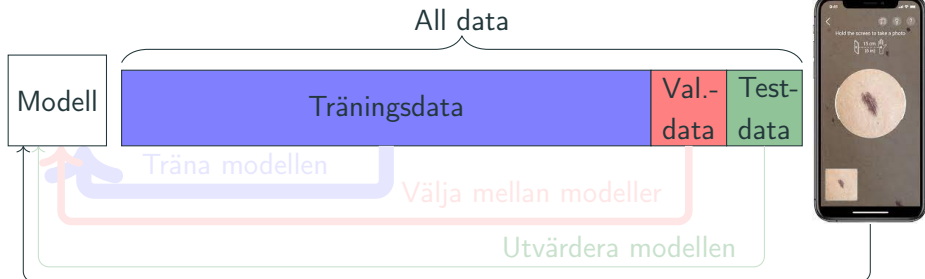
- **Träningsdata:** Används för att **träna** modellen
- **Valideringsdata:** Används för att **välja mellan** modeller

Tränings-, validerings-, och testdata



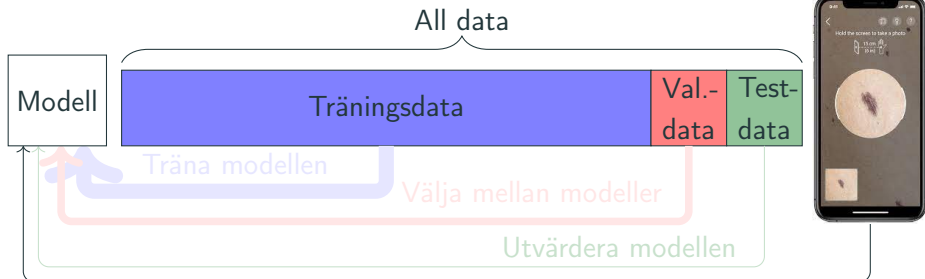
- **Träningsdata:** Används för att **träna** modellen
- **Valideringsdata:** Används för att **välja mellan** modeller
- **Testdata:** Används för att **utvärdera** modellen

Tränings-, validerings-, och testdata



- **Träningsdata**: Används för att **träna** modellen
- **Valideringsdata**: Används för att **välja mellan** modeller
- **Testdata**: Används för att **utvärdera** modellen

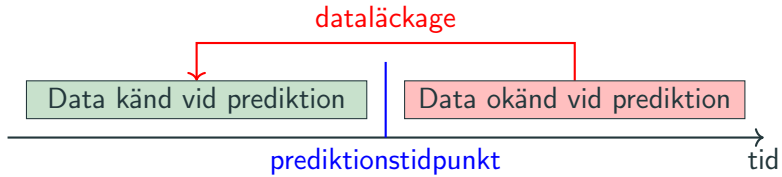
Tränings-, validerings-, och testdata



- **Träningsdata**: Används för att **träna** modellen
- **Valideringsdata**: Används för att **välja mellan** modeller
- **Testdata**: Används för att **utvärdera** modellen

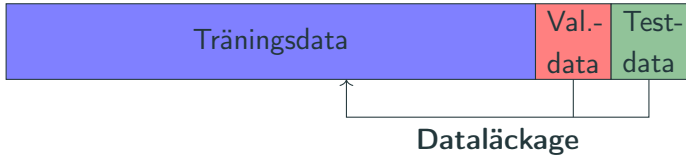
Validerings-/testdata ska motsvara verkligheten!

Fallgrop 1: Dataläckage (I/II)



Fallgrupp 1: Dataläckage (II/II)

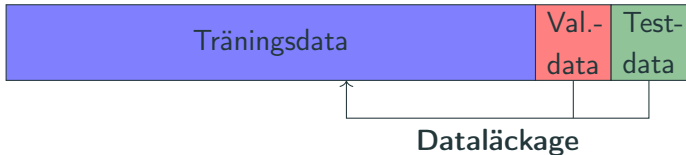
Hur kan dataläckage uppstå?



- Dataläckage: Från validerings-/test- till träningsdata.

Fallgrupp 1: Dataläckage (II/II)

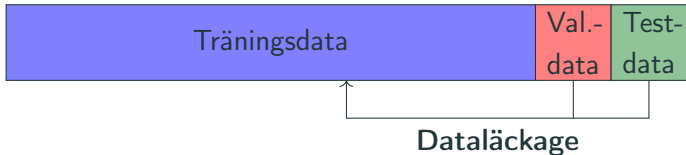
Hur kan dataläckage uppstå?



- **Dataläckage:** Från validerings-/test- till träningsdata.
 - *Dubletter i dataset:* Kopia av samma datapunkt finns i validerings-/testdata *och* träningsdata.

Fallgrupp 1: Dataläckage (II/II)

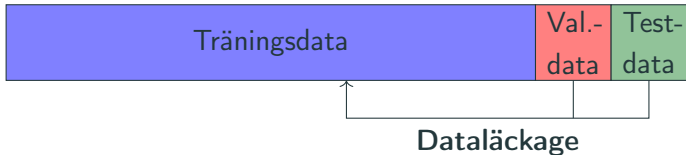
Hur kan dataläckage uppstå?



- **Dataläckage:** Från validerings-/test- till träningsdata.
 - *Dubletter i dataset:* Kopia av samma datapunkt finns i validerings-/testdata *och* träningsdata.
 - *Gemensam förbehandling av tränings-/validerings-/testdata*

Fallgrupp 1: Dataläckage (II/II)

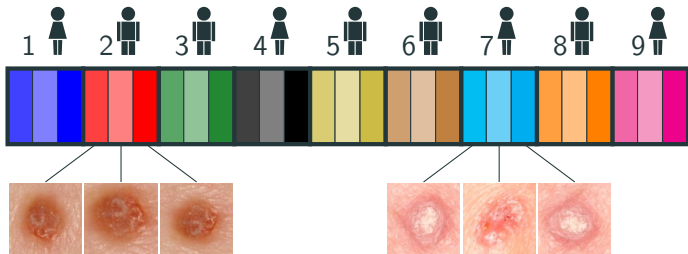
Hur kan dataläckage uppstå?



- **Dataläckage:** Från validerings-/test- till träningsdata.
 - *Dubletter i dataset:* Kopia av samma datapunkt finns i validerings-/testdata och träningsdata.
 - *Gemensam förbehandling av tränings-/validerings-/testdata*
 - *Gruppläckage*

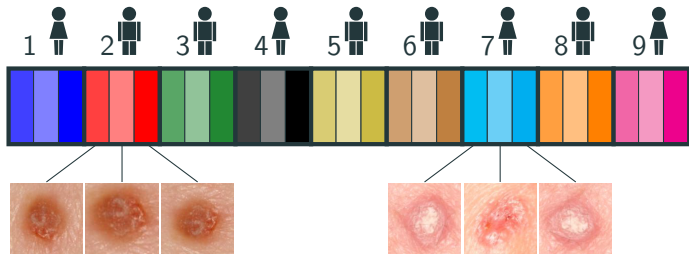
Ex) Gruppläckage

1) Fler bilder från samma individ

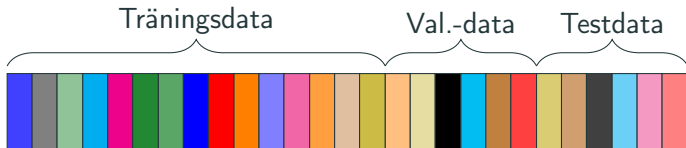


Ex) Gruppläckage

1) Fler bilder från samma individ

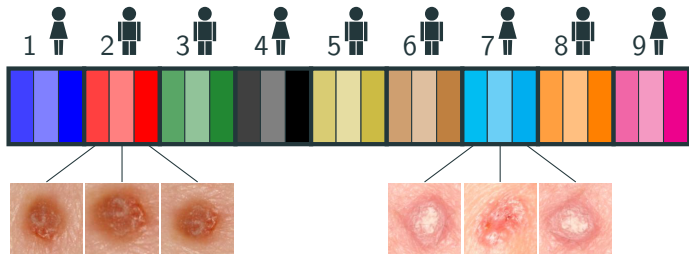


2) Blanda bilder och dela upp i träning/val/test.

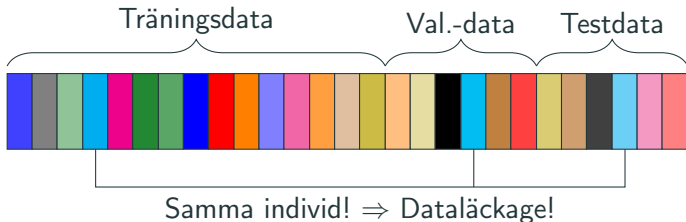


Ex) Gruppläckage

1) Fler bilder från samma individ

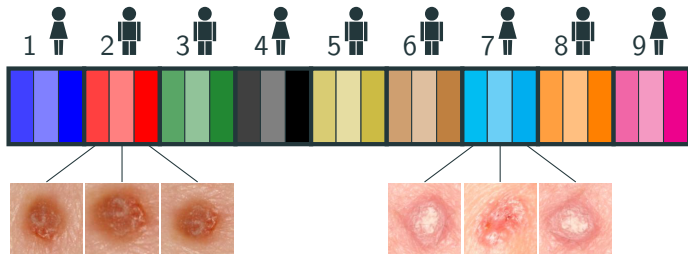


2) Blanda bilder och dela upp i träning/val/test.

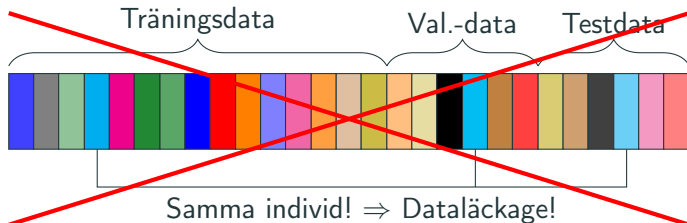


Ex) Gruppläckage

1) Fler bilder från samma individ

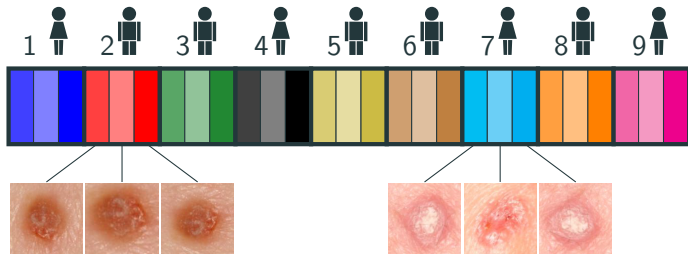


2) Blanda bilder och dela upp i träning/val/test.

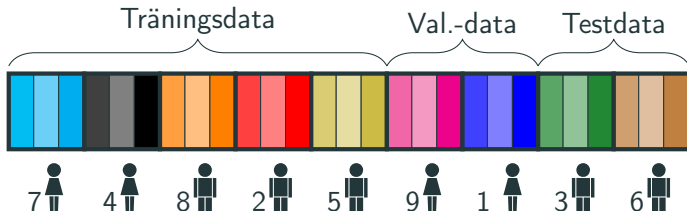


Ex) Gruppläckage

1) Fler bilder från samma individ

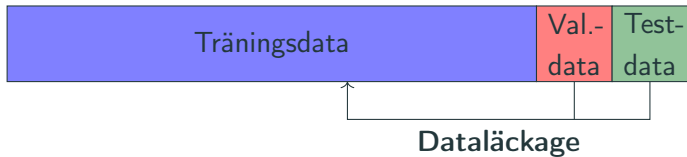


2) Blanda *individer* och dela upp i träning/val/test.



Fallgrop 1: Dataläckage (II/II)

Hur kan dataläckage uppstå?



- **Dataläckage:** Från validerings-/test- till träningsdata.
 - *Dubletter i dataset:* Kopia av samma datapunkt finns i validerings-/testdata *och* träningsdata.
 - *Gemensam förbehandling av tränings-/validerings-/testdata*
 - *Gruppläckage*



Sayash Kapoor & Arvind Narayanan. **Leakage and the reproducibility crisis in machine-learning-based science.** *Patterns*, 4(9), 2023.

Mål: Introducera ett strukturerat sätt att **utvärdera** och **förbättra** en maskininlärningsmodell.

1. Tränings-, validerings- och testdata

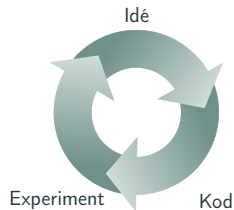
- Fallgrop 1: Dataläckage

2. Utvärderingsmättet

- Fallgrop 2: Inte ha *ett* utvärderingsmått

3. Träningsfel och generaliseringsfel

- Fallgrop 3: Att lägga tid på fel saker.



Utvärderingsmättet (I/III)

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Utvärderingsmättet (I/III)

Ex: God- och elakartade leverfläckar.
 $y = 0$: Godartad, $y = 1$: Elakartad.



	$y = 0$	$y = 1$	
<i>total</i>	95	5	100

Utvärderingsmättet (I/III)

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Modell A	$y = 0$	$y = 1$	<i>total</i>
$\hat{y}(\mathbf{x}) = 0$	85	1	86
$\hat{y}(\mathbf{x}) = 1$	10	4	14
<i>total</i>	95	5	100

Utvärderingsmättet (I/III)

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Modell A	$y = 0$	$y = 1$	<i>total</i>
$\hat{y}(\mathbf{x}) = 0$	85	1	86
$\hat{y}(\mathbf{x}) = 1$	10	4	14
<i>total</i>	95	5	100

$$\text{noggrannhet} = \frac{85 + 4}{100} = 89\%$$

Utvärderingsmättet (I/III)

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Modell A	$y = 0$	$y = 1$	<i>total</i>
$\hat{y}(\mathbf{x}) = 0$	85	1	86
$\hat{y}(\mathbf{x}) = 1$	10	4	14
<i>total</i>	95	5	100

Modell B	$y = 0$	$y = 1$	<i>total</i>
$\hat{y}(\mathbf{x}) = 0$	95	5	100
$\hat{y}(\mathbf{x}) = 1$	0	0	0
<i>total</i>	95	5	100

$$\text{noggrannhet} = \frac{85 + 4}{100} = 89\%$$

$$\text{noggrannhet} = \frac{95 + 0}{100} = 95\%$$

Utvärderingsmättet (I/III)

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Modell A	$y = 0$	$y = 1$	total
$\hat{y}(\mathbf{x}) = 0$	85	1	86
$\hat{y}(\mathbf{x}) = 1$	10	4	14
total	95	5	100

Modell B	$y = 0$	$y = 1$	total
$\hat{y}(\mathbf{x}) = 0$	95	5	100
$\hat{y}(\mathbf{x}) = 1$	0	0	0
total	95	5	100

$$\text{noggrannhet} = \frac{85 + 4}{100} = 89\%$$

$$\text{noggrannhet} = \frac{95 + 0}{100} = 95\%$$

Men är verkligen B "bättre" än A?

Utvärderingsmättet (I/III)

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Modell A	$y = 0$	$y = 1$	total
$\hat{y}(\mathbf{x}) = 0$	85	1	86
$\hat{y}(\mathbf{x}) = 1$	10	4	14
total	95	5	100

Modell B	$y = 0$	$y = 1$	total
$\hat{y}(\mathbf{x}) = 0$	95	5	100
$\hat{y}(\mathbf{x}) = 1$	0	0	0
total	95	5	100

$$\text{noggrannhet} = \frac{85 + 4}{100} = 89\%$$

$$\text{noggrannhet} = \frac{95 + 0}{100} = 95\%$$

Men är verkligen B "bättre" än A?

Modell B predikterar alltid $\hat{y}(\mathbf{x}) = 0$. Inte så användbar.

Utvärderingsmättet (II/III) - sensitivitet/specificitet

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Modell A	$y = 0$	$y = 1$	<i>total</i>
$\hat{y}(\mathbf{x}) = 0$	85	1	86
$\hat{y}(\mathbf{x}) = 1$	10	4	14
<i>total</i>	95	5	100

<i>specificitet</i>	<i>sensitivitet</i>
$= \frac{85}{95} = 89\%$	$= \frac{4}{5} = 80\%$

Utvärderingsmättet (II/III) - sensitivitet/specificitet

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Modell A	$y = 0$	$y = 1$	<i>total</i>
$\hat{y}(x) = 0$	85	1	86
$\hat{y}(x) = 1$	10	4	14
<i>total</i>	95	5	100

<i>specificitet</i>	<i>sensitivitet</i>
$= \frac{85}{95} = 89\%$	$= \frac{4}{5} = 80\%$

Modell B	$y = 0$	$y = 1$	<i>total</i>
$\hat{y}(x) = 0$	95	5	100
$\hat{y}(x) = 1$	0	0	0
<i>total</i>	95	5	100

<i>specificitet</i>	<i>sensitivitet</i>
$= \frac{95}{95} = 100\%$	$= \frac{0}{5} = 0\%$

Utvärderingsmättet (II/III) - sensitivitet/specificitet

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Modell A	$y = 0$	$y = 1$	total
$\hat{y}(\mathbf{x}) = 0$	85	1	86
$\hat{y}(\mathbf{x}) = 1$	10	4	14
total	95	5	100

<i>specificitet</i>	<i>sensitivitet</i>
$= \frac{85}{95} = 89\%$	$= \frac{4}{5} = 80\%$

Modell B	$y = 0$	$y = 1$	total
$\hat{y}(\mathbf{x}) = 0$	95	5	100
$\hat{y}(\mathbf{x}) = 1$	0	0	0
total	95	5	100

<i>specificitet</i>	<i>sensitivitet</i>
$= \frac{95}{95} = 100\%$	$= \frac{0}{5} = 0\%$

Verkar troligt att **A** är bättre...

Utvärderingsmättet (II/III) - sensitivitet/specifitet

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Modell A	$y = 0$	$y = 1$	<i>total</i>
$\hat{y}(\mathbf{x}) = 0$	85	1	86
$\hat{y}(\mathbf{x}) = 1$	10	4	14
<i>total</i>	95	5	100

<i>specificitet</i>	<i>sensitivitet</i>
$= \frac{85}{95} = 89\%$	$= \frac{4}{5} = 80\%$

Modell B	$y = 0$	$y = 1$	<i>total</i>
$\hat{y}(\mathbf{x}) = 0$	95	5	100
$\hat{y}(\mathbf{x}) = 1$	0	0	0
<i>total</i>	95	5	100

<i>specificitet</i>	<i>sensitivitet</i>
$= \frac{95}{95} = 100\%$	$= \frac{0}{5} = 0\%$

Verkar troligt att **A** är bättre...

Hur avgör vi generellt om **A** eller **B** är bättre?

Utvärderingsmättet (II/III) - sensitivitet/specificitet

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Modell A	$y = 0$	$y = 1$	total
$\hat{y}(x) = 0$	85	1	86
$\hat{y}(x) = 1$	10	4	14
total	95	5	100

Modell B	$y = 0$	$y = 1$	total
$\hat{y}(x) = 0$	95	5	100
$\hat{y}(x) = 1$	0	0	0
total	95	5	100

<i>specificitet</i>	<i>sensitivitet</i>
$= \frac{85}{95} = 89\%$	$= \frac{4}{5} = 80\%$

<i>specificitet</i>	<i>sensitivitet</i>
$= \frac{95}{95} = 100\%$	$= \frac{0}{5} = 0\%$

Verkar troligt att **A** är bättre...

Hur avgör vi generellt om **A** eller **B** är bättre?

Ett förslag: *Balanserad noggrannhet* = $\frac{\text{specificitet} + \text{sensitivitet}}{2}$

Utvärderingsmåttet (III/III)

Vilket utvärderingsmått du ska välja beror på ditt problem!

Utvärderingsmåttet (III/III)

Vilket utvärderingsmått du ska välja beror på ditt problem!

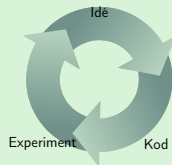
Fallgrop 2: Att inte bestämt sig för ett utvärderingsmått.

Utvärderingsmålet (III/III)

Vilket utvärderingsmål du ska välja beror på ditt problem!

Fallgrop 2: Att inte bestämt sig för **ett** utvärderingsmål.

Valideringsdata + **ett** utvärderingsmål
definierar ditt maskininlärningsproblem och
snabbar sökandet av bättre modeller.

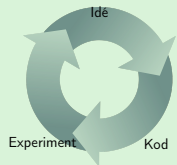


Utvärderingsmåttet (III/III)

Vilket utvärderingsmått du ska välja beror på ditt problem!

Fallgrupp 2: Att inte bestämt sig för **ett** utvärderingsmått.

Valideringsdata + **ett** utvärderingsmått
definierar ditt maskininlärningsproblem och
snabbar sökandet av bättre modeller.



Tillvägagångssätt:

1. Definiera ditt problem (valideringsdata + utvärderingsmått).
2. Anpass din modell så du blir bra på att lösa detta problem.

Mål: Introducera ett strukturerat sätt att **utvärdera** och **förbättra** en maskininlärningsmodell.

1. Tränings-, validerings- och testdata

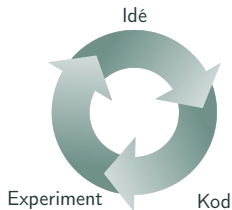
- Fallgrop 1: Dataläckage

2. Utvärderingsmättet

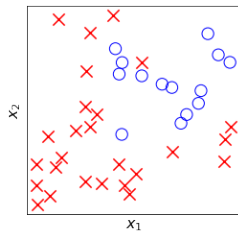
- Fallgrop 2: Inte ha *ett* utvärderingsmått

3. Träningsfel och generaliseringsfel

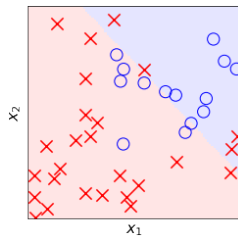
- Fallgrop 3: Att lägga tid på fel saker.



Träningsdata

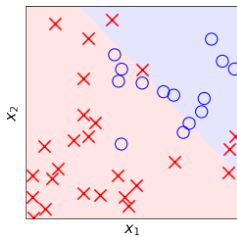


Träningsdata

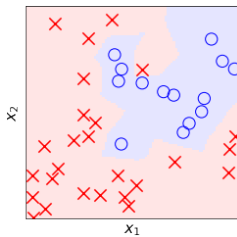


Underanpassning

Träningsdata



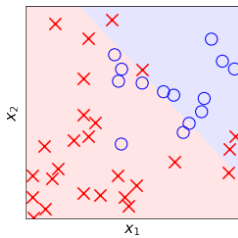
Underanpassning



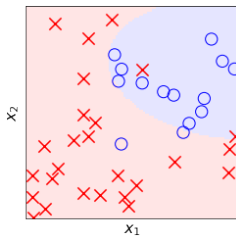
Överanpassning

Träningsfel och generaliseringsfel (I/III)

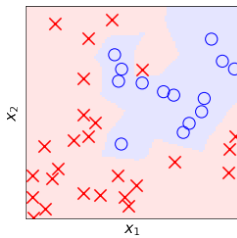
Träningsdata



Underanpassning



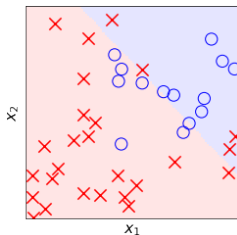
"Precis rätt"



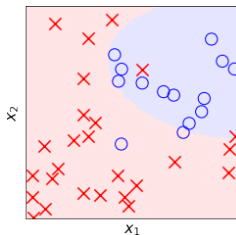
Överanpassning

Träningsfel och generaliseringsfel (I/III)

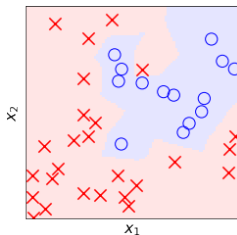
Träningsdata



Underanpassning
Högt träningsfel



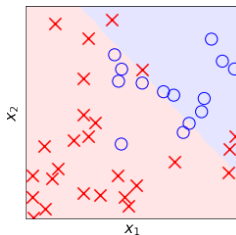
"Precis rätt"



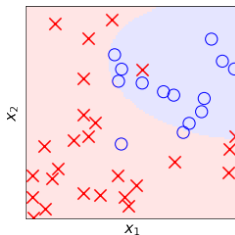
Överanpassning
Lågt träningsfel

Träningsfel och generaliseringsfel (I/III)

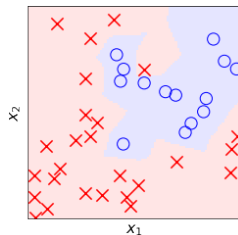
Träningsdata



Underanpassning
Högt träningsfel



"Precis rätt"



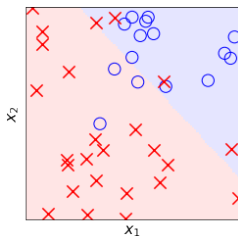
Överanpassning
Lågt träningsfel

Vi vill minimera felet på valideringsdata!

$$\text{valideringsfel} = \text{träningsfel} + \underbrace{(\text{valideringsfel} - \text{träningsfel})}_{\text{generaliseringsfel}}$$

Träningsfel och generaliseringsfel (I/III)

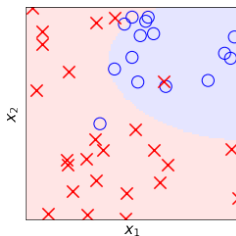
Valideringsdata



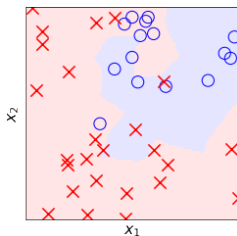
Underanpassning

Högt träningsfel

Lågt generaliseringsfel



"Precis rätt"



Överanpassning

Lågt träningsfel

Högt generaliseringsfel

Vi vill minimera felet på valideringsdata!

$$\text{valideringsfel} = \text{träningsfel} + \underbrace{(\text{valideringsfel} - \text{träningsfel})}_{\text{generaliseringsfel}}$$

Träningsfel och generaliseringsfel (II/III)

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Träningsfel och generaliseringsfel (II/III)

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Träningsfel	Valideringsfel	
9%	10%	Högt träningsfel, lågt generaliseringsfel

$$\text{valideringsfel} = \text{träningsfel} + \underbrace{(\text{valideringsfel} - \text{träningsfel})}_{\text{generaliseringsfel}}$$

Träningsfel och generaliseringsfel (II/III)

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Träningsfel	Valideringsfel	
9%	10%	Högt träningsfel, lågt generaliseringsfel
1%	8%	Lågt träningsfel, högt generaliseringsfel

$$\text{valideringsfel} = \text{träningsfel} + \underbrace{(\text{valideringsfel} - \text{träningsfel})}_{\text{generaliseringsfel}}$$

Träningsfel och generaliseringsfel (II/III)

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



Träningsfel	Valideringsfel	
9%	10%	Högt träningsfel, lågt generaliseringsfel
1%	8%	Lågt träningsfel, högt generaliseringsfel
1%	2%	Lågt träningsfel, lågt generaliseringsfel

$$\text{valideringsfel} = \text{träningsfel} + \underbrace{(\text{valideringsfel} - \text{träningsfel})}_{\text{generaliseringsfel}}$$

Träningsfel och generaliseringsfel (II/III)

Ex: God- och elakartade leverfläckar.

$y = 0$: Godartad, $y = 1$: Elakartad.



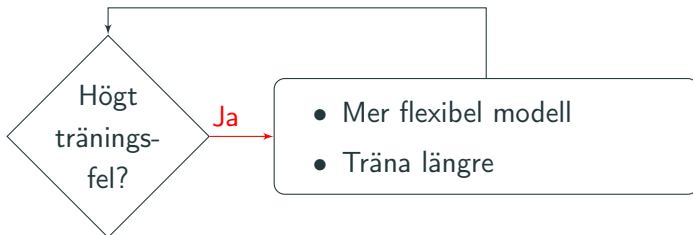
Träningsfel	Valideringsfel	
9%	10%	Högt träningsfel, lågt generaliseringsfel
1%	8%	Lågt träningsfel, högt generaliseringsfel
1%	2%	Lågt träningsfel, lågt generaliseringsfel
10%	20%	Högt träningsfel, högt generaliseringsfel

$$\text{valideringsfel} = \text{träningsfel} + \underbrace{(\text{valideringsfel} - \text{träningsfel})}_{\text{generaliseringsfel}}$$

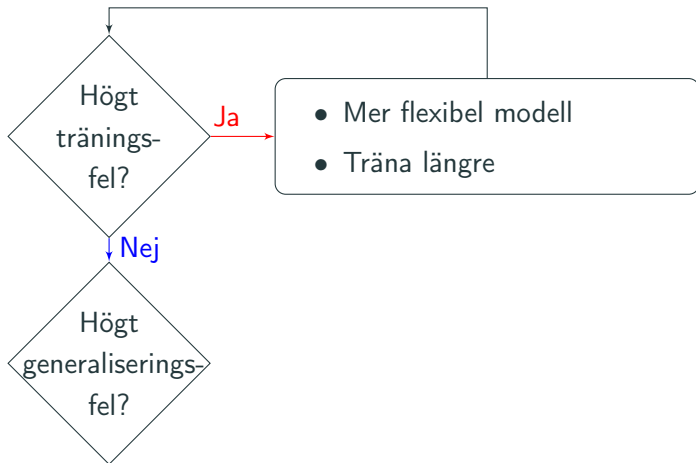
Enkelt recept för maskininlärning



Enkelt recept för maskininlärning

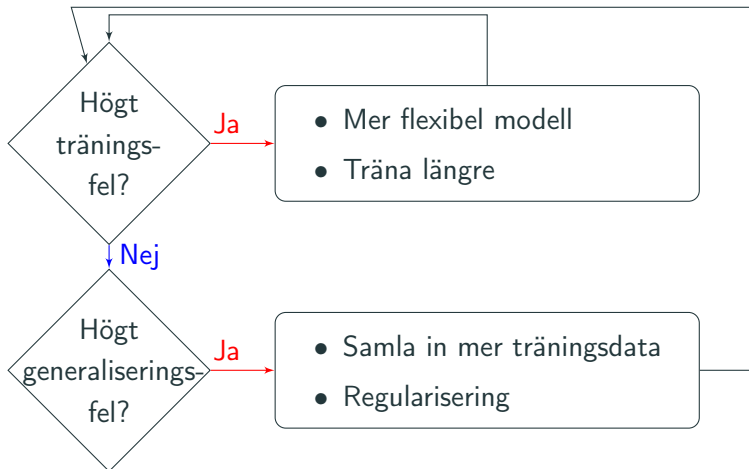


Enkelt recept för maskininlärning



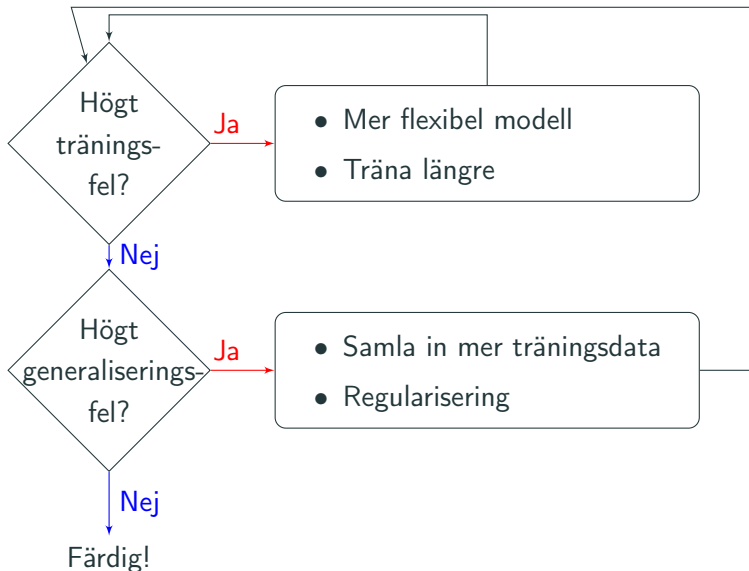
Träningsfel och generaliseringsfel (III/III)

Enkelt recept för maskininlärning



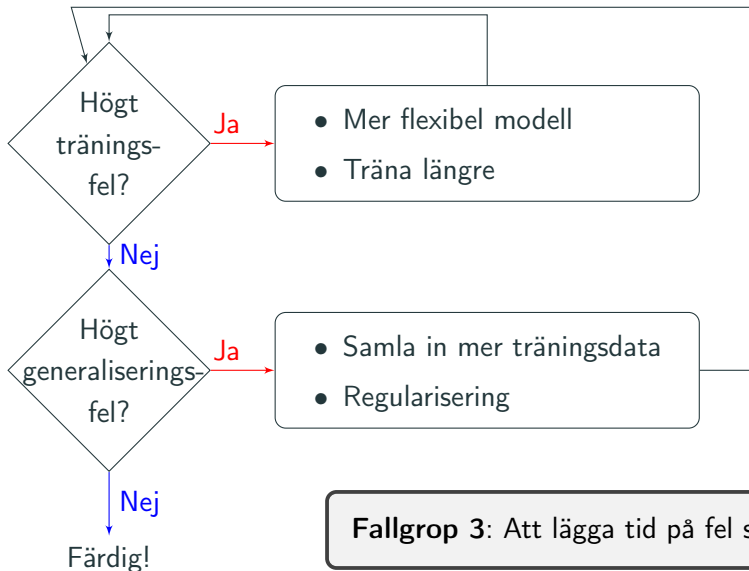
Träningsfel och generaliseringsfel (III/III)

Enkelt recept för maskininlärning

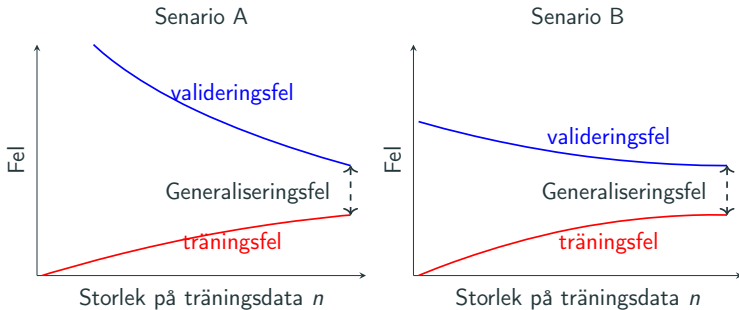


Träningsfel och generaliseringsfel (III/III)

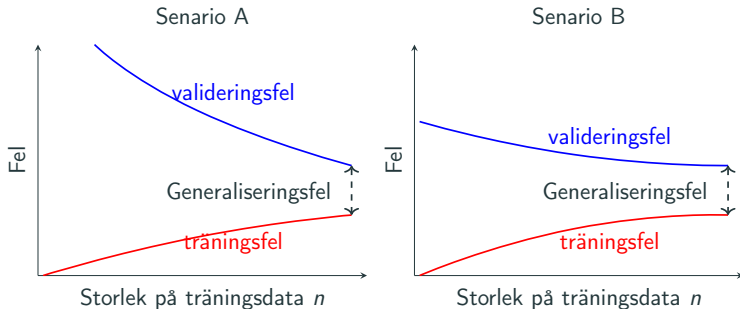
Enkelt recept för maskininlärning



Hur mycket kan modellen förbättras med mer data?



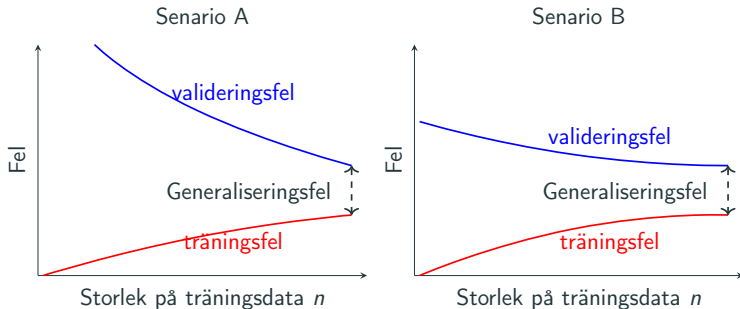
Hur mycket kan modellen förbättras med mer data?



Finns också många bra alternativ till att samla in mer data:

- Dataaugmentering

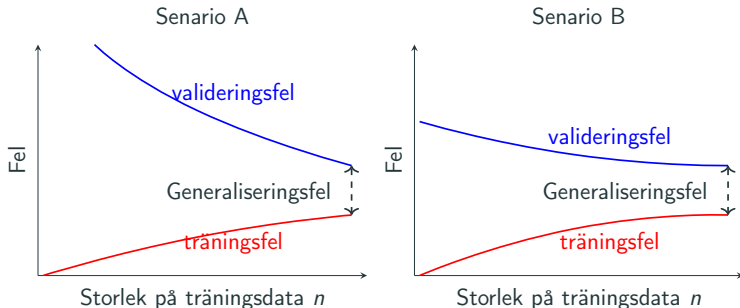
Hur mycket kan modellen förbättras med mer data?



Finns också många bra alternativ till att samla in mer data:

- Dataaugmentering
- Överföringsinlärning (*eng.* transfer learning)

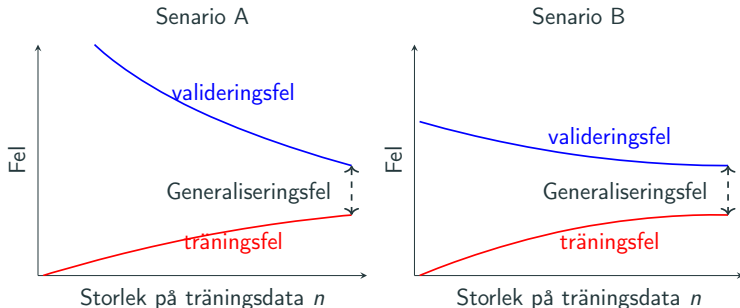
Hur mycket kan modellen förbättras med mer data?



Finns också många bra alternativ till att samla in mer data:

- Dataaugmentering
- Överföringsinlärning (*eng.* transfer learning)
- Självövervakad inlärning (*eng.* self-supervised learning)

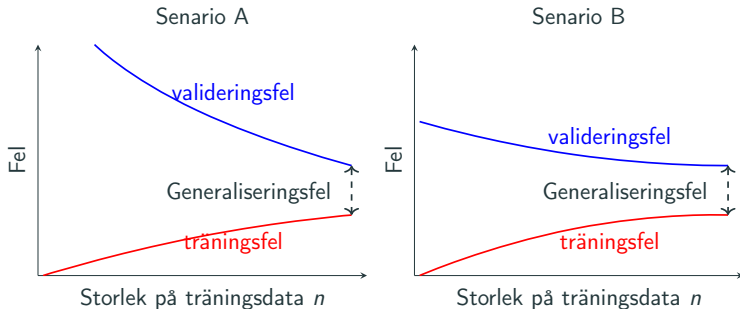
Hur mycket kan modellen förbättras med mer data?



Finns också många bra alternativ till att samla in mer data:

- Dataaugmentering
- Överföringsinlärning (*eng.* transfer learning)
- Självövervakad inlärning (*eng.* self-supervised learning)

Hur mycket kan modellen förbättras med mer data?



Finns också många bra alternativ till att samla in mer data:

- Dataaugmentering
- Överföringsinlärning (*eng.* transfer learning)
- Självövervakad inlärning (*eng.* self-supervised learning)

Men det är en annan föreläsning...

Mål: Introducera ett strukturerat sätt att **utvärdera** och **förbättra** en maskininlärningsmodell.

1. Dela upp data i tränings-, validerings- och testdata

Mål: Introducera ett strukturerat sätt att **utvärdera** och **förbättra** en maskininlärningsmodell.

1. Dela upp data i tränings-, validerings- och testdata
 - Se till att validering/test motsvarar verkligheten!

Mål: Introducera ett strukturerat sätt att **utvärdera** och **förbättra** en maskininlärningsmodell.

1. Dela upp data i tränings-, validerings- och testdata
 - Se till att validering/test motsvarar verkligheten!
 - Se upp för dataläckage

Mål: Introducera ett strukturerat sätt att **utvärdera** och **förbättra** en maskininlärningsmodell.

1. Dela upp data i tränings-, validerings- och testdata
 - Se till att validering/test motsvarar verkligheten!
 - Se upp för dataläckage
2. Välj **ett** utvärderingsmåttet

Mål: Introducera ett strukturerat sätt att **utvärdera** och **förbättra** en maskininlärningsmodell.

1. Dela upp data i tränings-, validerings- och testdata
 - Se till att validering/test motsvarar verkligheten!
 - Se upp för dataläckage
2. Välj **ett** utvärderingsmåttet
 - Valideringsdata+utvärderingsmåttet definierar ML-problemet

Mål: Introducera ett strukturerat sätt att **utvärdera** och **förbättra** en maskininlärningsmodell.

1. Dela upp data i tränings-, validerings- och testdata
 - Se till att validering/test motsvarar verkligheten!
 - Se upp för dataläckage
2. Välj **ett** utvärderingsmåttet
 - Valideringsdata+utvärderingsmåttet definierar ML-problemet
3. Förbättra modellen systematiskt

Mål: Introducera ett strukturerat sätt att **utvärdera** och **förbättra** en maskininlärningsmodell.

1. Dela upp data i tränings-, validerings- och testdata
 - Se till att validering/test motsvarar verkligheten!
 - Se upp för dataläckage
2. Välj **ett** utvärderingsmåttet
 - Valideringsdata+utvärderingsmåttet definierar ML-problemet
3. Förbättra modellen systematiskt
 - Börja med en enkel modell.

Mål: Introducera ett strukturerat sätt att **utvärdera** och **förbättra** en maskininlärningsmodell.

1. Dela upp data i tränings-, validerings- och testdata
 - Se till att validering/test motsvarar verkligheten!
 - Se upp för dataläckage
2. Välj **ett** utvärderingsmåttet
 - Valideringsdata+utvärderingsmåttet definierar ML-problemet
3. Förbättra modellen systematiskt
 - Börja med en enkel modell.
 - Använd tränings- och generaliseringsfelet som vägledning

Mål: Introducera ett strukturerat sätt att **utvärdera** och **förbättra** en maskininlärningsmodell.

1. Dela upp data i tränings-, validerings- och testdata
 - Se till att validering/test motsvarar verkligheten!
 - Se upp för dataläckage
2. Välj **ett** utvärderingsmåttet
 - Valideringsdata+utvärderingsmåttet definierar ML-problemet
3. Förbättra modellen systematiskt
 - Börja med en enkel modell.
 - Använd tränings- och generaliseringsfelet som vägledning

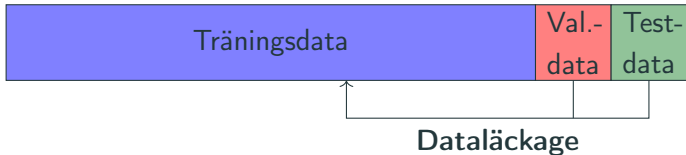
Mål: Introducera ett strukturerat sätt att **utvärdera** och **förbättra** en maskininlärningsmodell.

1. Dela upp data i tränings-, validerings- och testdata
 - Se till att validering/test motsvarar verkligheten!
 - Se upp för dataläckage
2. Välj **ett** utvärderingsmåttet
 - Valideringsdata+utvärderingsmåttet definierar ML-problemet
3. Förbättra modellen systematiskt
 - Börja med en enkel modell.
 - Använd tränings- och generaliseringsfelet som vägledning

Lycka till med ditt maskininlärningsprojekt!

Fallgrupp 1: Dataläckage (II/II)

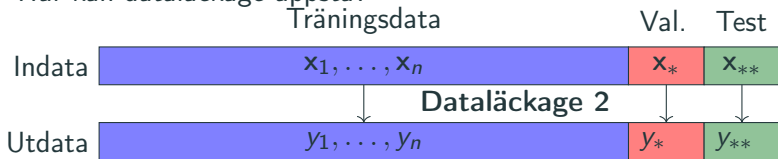
Hur kan dataläckage uppstå?



- **Dataläckage:** Från validerings-/test- till träningsdata.
 - *Dubletter i dataset:* Kopia av samma datapunkt finns i validerings-/testdata och träningsdata.
 - *Gemensam förbehandling av tränings-/validerings-/testdata*
 - *Gruppläckage*

Fallgrop 1: Dataläckage (II/II)

Hur kan dataläckage uppstå?



- **Dataläckage:** Från validerings-/test- till träningsdata.
 - *Dubletter i dataset:* Kopia av samma datapunkt finns i validerings-/testdata och träningsdata.
 - *Gemensam förbehandling av tränings-/validerings-/testdata*
 - *Gruppläckage*
- **Dataläckage 2:** Från indata x till utdata y .

Ex) Dataläckage från indata x till utdata y

Example 1:

Indata x				Utdata y
EmployeeID	Title	ExperienceYears	MonthlySalaryGBP	AnnualIncomeUSD
315981	Data Scientist	3	5,000.00	78,895.44
4691	Data Scientist	4	5,500.00	86,784.98
23598	Data Scientist	5	6,200.00	97,830.35

Ex) Dataläckage från indata x till utdata y

Example 1: Utdata funktion av en annan kolumn

Indata x				Utdata y
EmployeeID	Title	ExperienceYears	MonthlySalaryGBP	AnnualIncomeUSD
315981	Data Scientist	3	5,000.00	78,895.44
4691	Data Scientist	4	5,500.00	86,784.98
23598	Data Scientist	5	6,200.00	97,830.35

- Data kan ha olika formatering eller enheter i olika kolumner.
- Glömma ta bort kopior kan introducera dataläckage.

Ex) Dataläckage från indata x till utdata y

Example 1: Utdata funktion av en annan kolumn

Indata x				Utdata y
EmployeeID	Title	ExperienceYears	MonthlySalaryGBP	AnnualIncomeUSD
315981	Data Scientist	3	5,000.00	78,895.44
4691	Data Scientist	4	5,500.00	86,784.98
23598	Data Scientist	5	6,200.00	97,830.35

- Data kan ha olika formatering eller enheter i olika kolumner.
- Glömma ta bort kopior kan introducera dataläckage.

Exempel 2:

Indata x					Utdata y
Education	Married	AnnualIncome	Purpose	LatePaymentReminders	IsBadLoan
1	Y	80k	Car Purchase	0	0
3	N	120k	Small Business	3	1
1	Y	85k	House Purchase	5	1
2	N	72k	Marriage	1	0

Ex) Dataläckage från indata x till utdata y

Example 1: Utdata funktion av en annan kolumn

Indata x				Utdata y
EmployeeID	Title	ExperienceYears	MonthlySalaryGBP	AnnualIncomeUSD
315981	Data Scientist	3	5,000.00	78,895.44
4691	Data Scientist	4	5,500.00	86,784.98
23598	Data Scientist	5	6,200.00	97,830.35

- Data kan ha olika formatering eller enheter i olika kolumner.
- Glömma ta bort kopior kan introducera dataläckage.

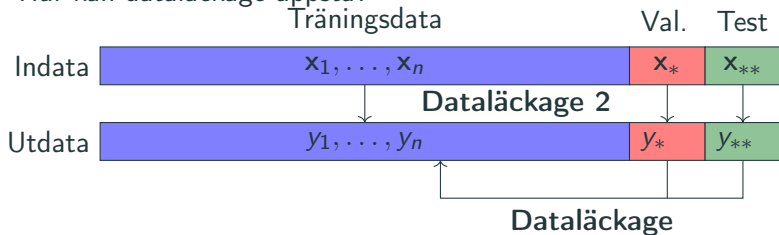
Exempel 2: Förändlig data pga avsaknad av ögonblicksbild

Indata x					Utdata y
Education	Married	AnnualIncome	Purpose	LatePaymentReminders	IsBadLoan
1	Y	80k	Car Purchase	0	0
3	N	120k	Small Business	3	1
1	Y	85k	House Purchase	5	1
2	N	72k	Marriage	1	0

- Databas blir överskriven när ny data kommer in.
- Denna senare info finns inte under vid prediktionsögonblicket

Fallgrop 1: Dataläckage (II/II)

Hur kan dataläckage uppstå?

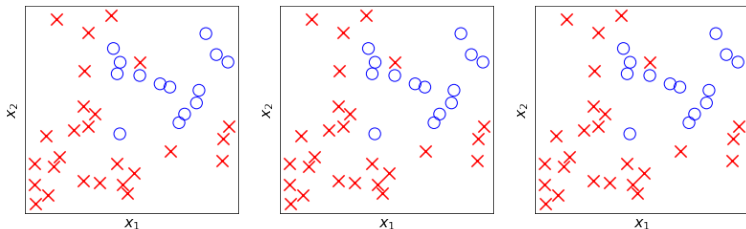


- **Dataläckage:** Från validerings-/test- till träningsdata.
 - *Dubletter i dataset:* Kopia av samma datapunkt finns i validerings-/testdata och träningsdata.
 - *Gemensam förbehandling av tränings-/validerings-/testdata*
 - *Gruppläckage*
- **Dataläckage 2:** Från indata x till utdata y .

Träningsfel och valideringsfel

- E_{train} : Träningsfel,
- E_{val} : Valideringsfel, Minimera $E_{\text{val}} = E_{\text{train}} + \underbrace{(E_{\text{val}} - E_{\text{train}})}_{\text{generaliseringsfel}}$

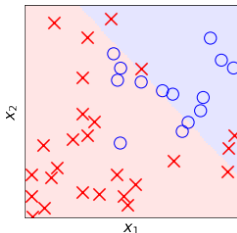
Träningsdata



Träningsfel och valideringsfel

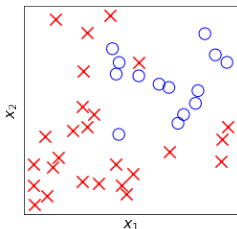
- E_{train} : Träningsfel,
- E_{val} : Valideringsfel, Minimera $E_{\text{val}} = E_{\text{train}} + \underbrace{(E_{\text{val}} - E_{\text{train}})}_{\text{generaliseringsfel}}$

Träningsdata

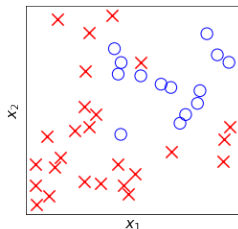


Högt träningsfel

$$(E_{\text{train}} = 5)$$



$$(E_{\text{train}} = 2)$$

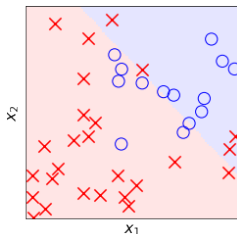


$$(E_{\text{train}} = 0)$$

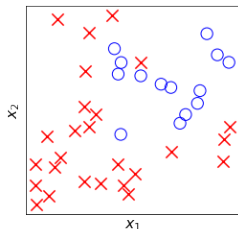
Träningsfel och valideringsfel

- E_{train} : Träningsfel,
- E_{val} : Valideringsfel, Minimera $E_{\text{val}} = E_{\text{train}} + \underbrace{(E_{\text{val}} - E_{\text{train}})}_{\text{generaliseringsfel}}$

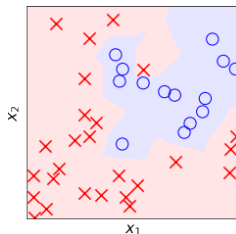
Träningsdata



Högt träningsfel
($E_{\text{train}} = 5$)



($E_{\text{train}} = 2$)

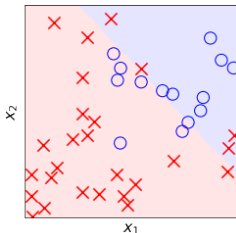


Lågt träningsfel
($E_{\text{train}} = 0$)

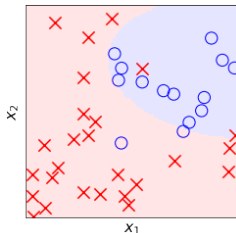
Träningsfel och valideringsfel

- E_{train} : Träningsfel,
- E_{val} : Valideringsfel, Minimera $E_{\text{val}} = E_{\text{train}} + \underbrace{(E_{\text{val}} - E_{\text{train}})}_{\text{generaliseringsfel}}$

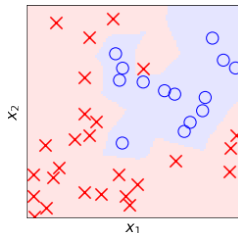
Träningsdata



Högt träningsfel
($E_{\text{train}} = 5$)



"Precis rätt"
($E_{\text{train}} = 2$)

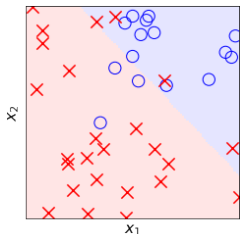


Lågt träningsfel
($E_{\text{train}} = 0$)

Träningsfel och valideringsfel

- E_{train} : Träningsfel,
- E_{val} : Valideringsfel, Minimera $E_{\text{val}} = E_{\text{train}} + \underbrace{(E_{\text{val}} - E_{\text{train}})}_{\text{generaliseringsfel}}$

Valideringsdata

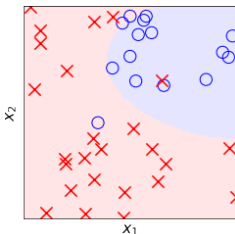


Högt träningsfel

$$(E_{\text{train}} = 5)$$

$$(E_{\text{val}} = 6)$$

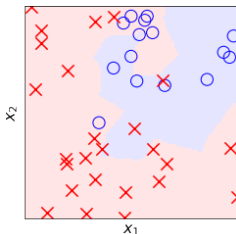
Lågt generaliseringsfel



"Precis rätt"

$$(E_{\text{train}} = 2)$$

$$(E_{\text{val}} = 4)$$



Lågt träningsfel

$$(E_{\text{train}} = 0)$$

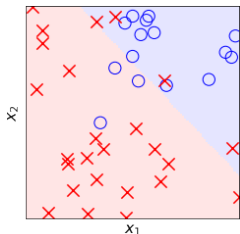
$$(E_{\text{val}} = 9)$$

Högt generaliseringsfel ^{26/22}

Träningsfel och valideringsfel

- E_{train} : Träningsfel,
- E_{val} : Valideringsfel, Minimera $E_{\text{val}} = E_{\text{train}} + \underbrace{(E_{\text{val}} - E_{\text{train}})}_{\text{generaliseringsfel}}$

Valideringsdata

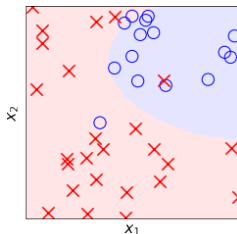


Högt träningsfel

$$(E_{\text{train}} = 5)$$

$$(E_{\text{val}} = 6)$$

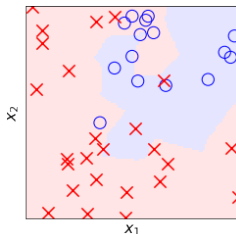
Lågt generaliseringsfel



"Precis rätt"

$$(E_{\text{train}} = 2)$$

$$(E_{\text{val}} = 4)$$



Lågt träningsfel

$$(E_{\text{train}} = 0)$$

$$(E_{\text{val}} = 9)$$

Högt generaliseringsfel ^{26/22}