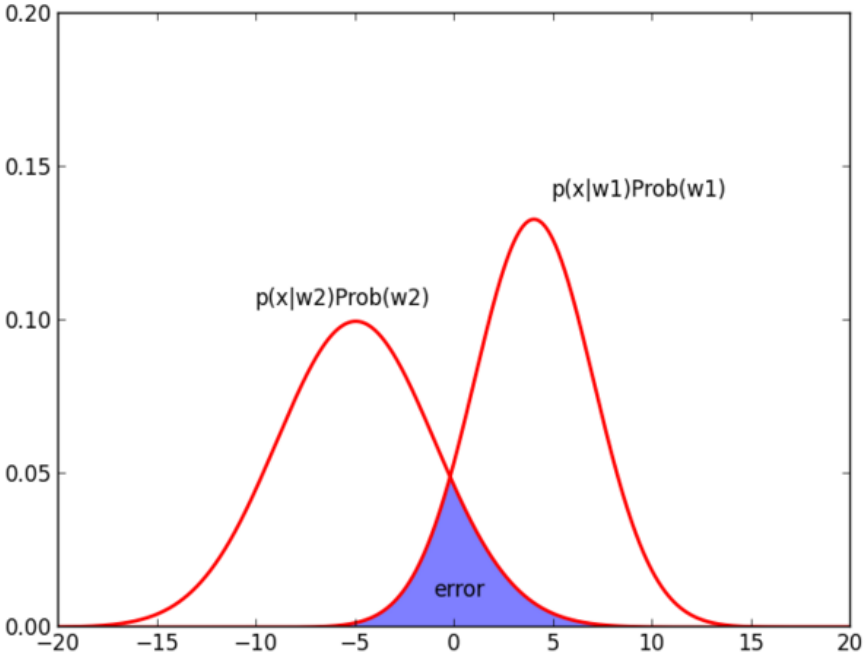
Probabilistiske modeller

Anvendt sandsynlighedsregning

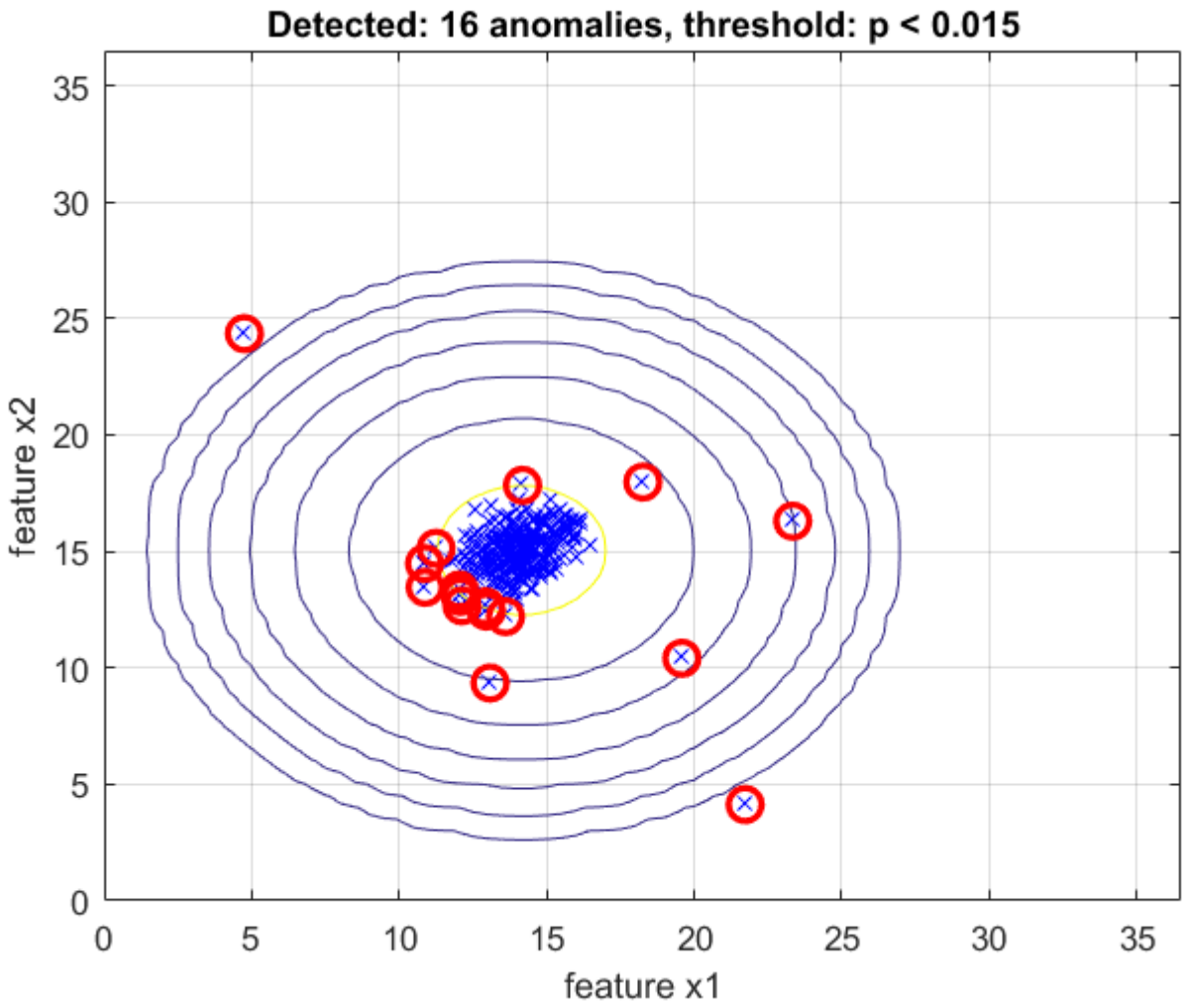
Fordele ved probabilistiske modeller



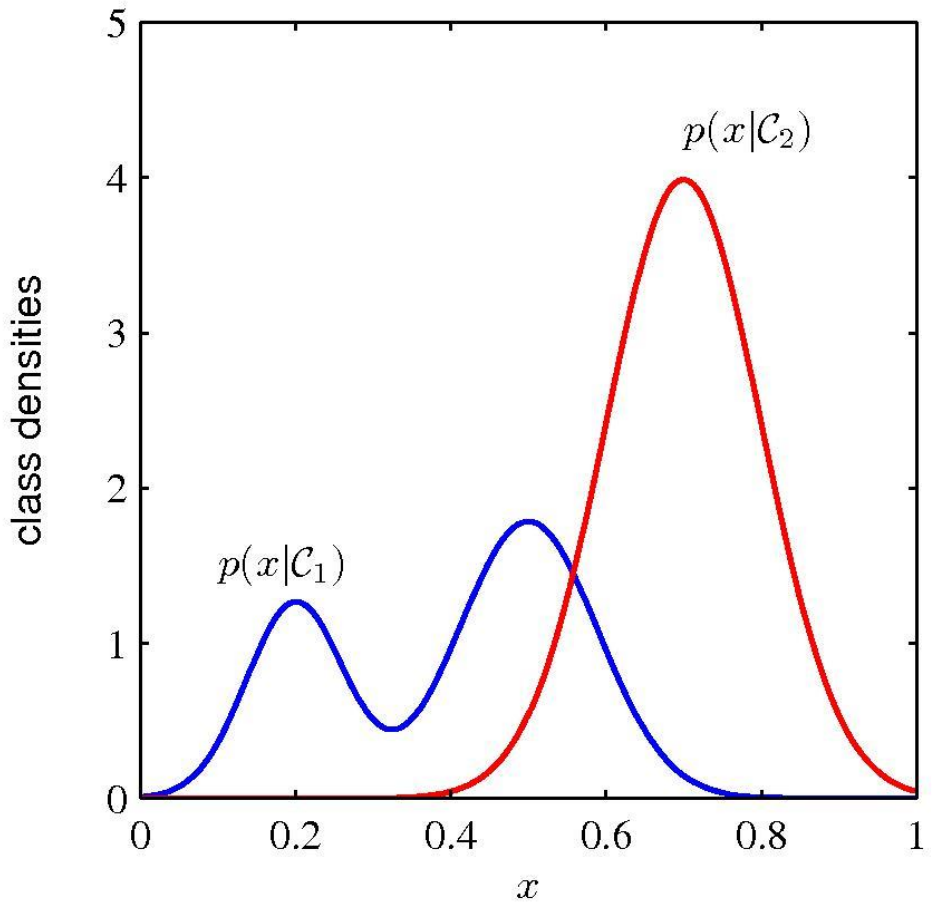
Basics of Decision Theory

	Surgery Performed	Surgery Not Performed
Cancer Present (C1)	**Very Good** Score: 100%	**Very Bad** Score: 0%
Cancer Absent (C2)	**Not Good** 40%	**Good** 85%

Fordele - Outlier detection



Sandsynlighedsfordelinger - klassifikation



Betingede sandsynligheder

Betinget sandsynlighed - Eksempel

Eksempel:
Køns-fordeling af arbejdsløse/ikke-arbejdsløse med studentereksamen i en lille by

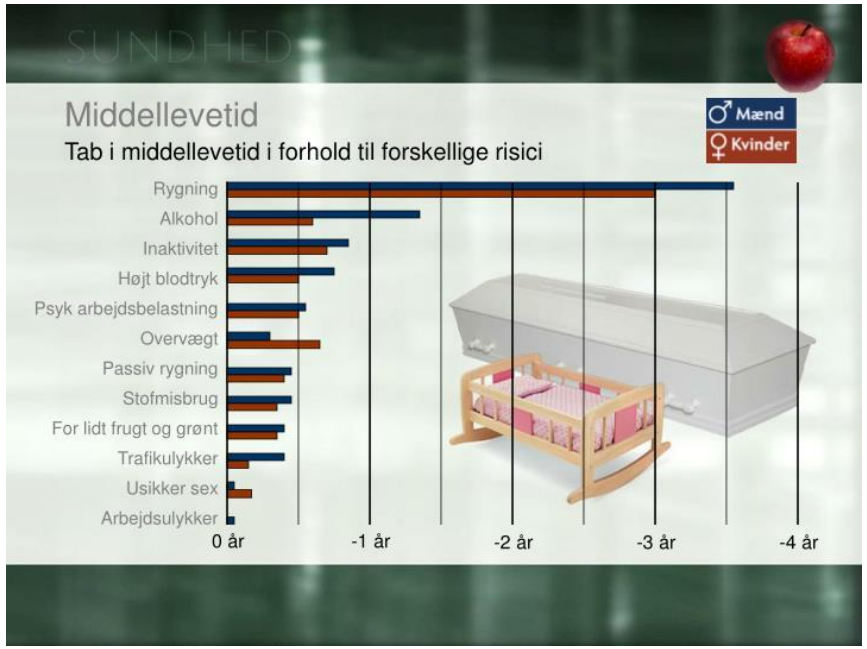
	I arbejde	Arbejdsløs	Total
Mand	460	40	500
Kvinde	140	260	400
Total	600	300	900

$$P(\text{mand} \mid \text{i arbejde}) = \frac{P(\text{mand \& i arbejde})}{P(\text{i arbejde})} = \frac{460/900}{600/900} = \frac{460}{600} = \frac{23}{30} = 76.7\%$$

$$P(\text{mand} \mid \text{arbejdsløs}) = \frac{P(\text{mand \& arbejdsløs})}{P(\text{arbejdsløs})} = \frac{40/900}{300/900} = \frac{40}{300} = \frac{2}{15} = 13.3\%$$

Tabel 1. Beregnet middellevetid for mænd og kvinder 1895–1900 (1961–65).

	Mænd	Kvinder	Kvinder i % af mænd
1895–1900	50.2 år	53.2 år	106
1936–1940	63.5 -	65.8 -	104
1941–1945	65.6 -	67.7 -	103
1946–1950	67.8 -	70.1 -	103
1951–1955	69.8 -	72.6 -	104
1956–1960	70.4 -	73.8 -	105
1961–1965	70.3 -	74.5 -	106



Bayes' theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X) = \sum_{Y} p(X|Y)p(Y)$$

posterior $\propto$ likelihood × prior

Bayes' theorem for classification

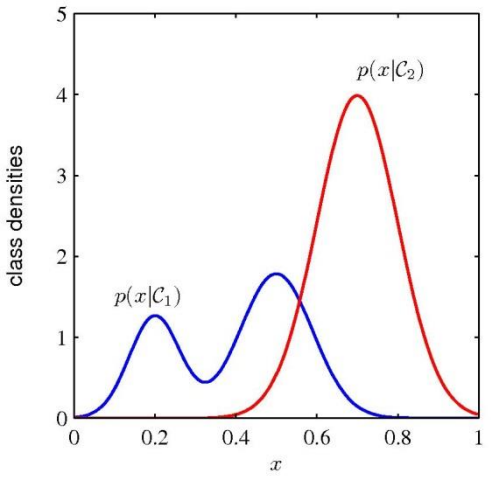
$$P(C|x) = \frac{P(x|C)P(C)}{P(x)} = \frac{P(x|C)P(C)}{\sum_{C=1}^{N_c} P(x|C)P(C)}$$

P(C|x) – Posterior class probability
P(C) – probability of (some) class C
P(x) – probability density of random variable x at value x (sloppy notation, but that's normal..)
N_c – Number of classes

NOTE: Arg $\max_C$ P(C|x) = Arg $\max_C$ P(x|C)P(C)
- and if all class have same prior probability (ie. P(C) uniformly distributed), then Arg $\max_C$ P(C|x) = Arg $\max_C$ P(x|C)

Classification using probabilistic models
- Generative models

1. Choose probabilistic model/distribution for p(x|C)
   - e.g. Gaussian distribution
2. (Sometimes) choose initial model parameters
   - Number of mixtures, hyperparameters,..
3. Infer/learn parameters of the model by maximizing the Log-likelihood function for each class separately – using a training data set
   - e.g. μ and σ for the Gaussian for each class
4. Find the posterior class probabilities P(C | x) using Bayes' theorem
5. Now, we're ready to use P(C | x) on a new (test) set.



LDA / QDA classifier i Scikit Learn

Both LDA and QDA can be derived from simple probabilistic models which model the class conditional distribution of the data $P(X|y=k)$ for each class k. Predictions can then be obtained by using Bayes' rule:

$$P(y=k|X) = \frac{P(X|y=k)P(y=k)}{P(X)} = \frac{P(X|y=k)P(y=k)}{\sum_l P(X|y=l) \cdot P(y=l)}$$

and we select the class k which maximizes this conditional probability.

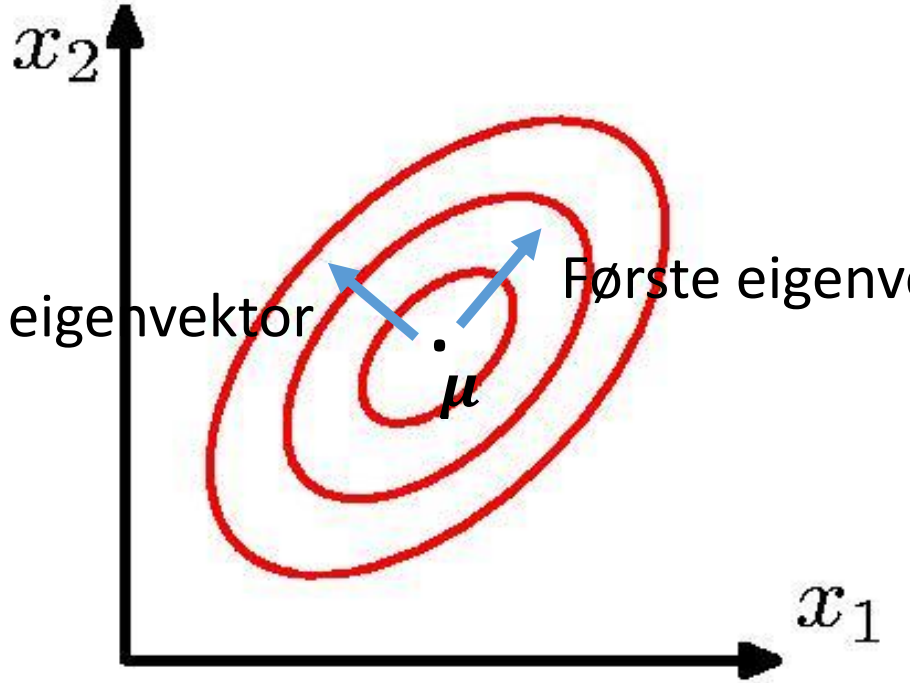
More specifically, for linear and quadratic discriminant analysis, $P(X|y)$ is modeled as a multivariate Gaussian distribution with density:

$$P(X|y=k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(X-\mu_k)^t \Sigma_k^{-1}(X-\mu_k)\right)$$

where d is the number of features.

Multivariate Gaussian
(ie. more than one dimension..)

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$



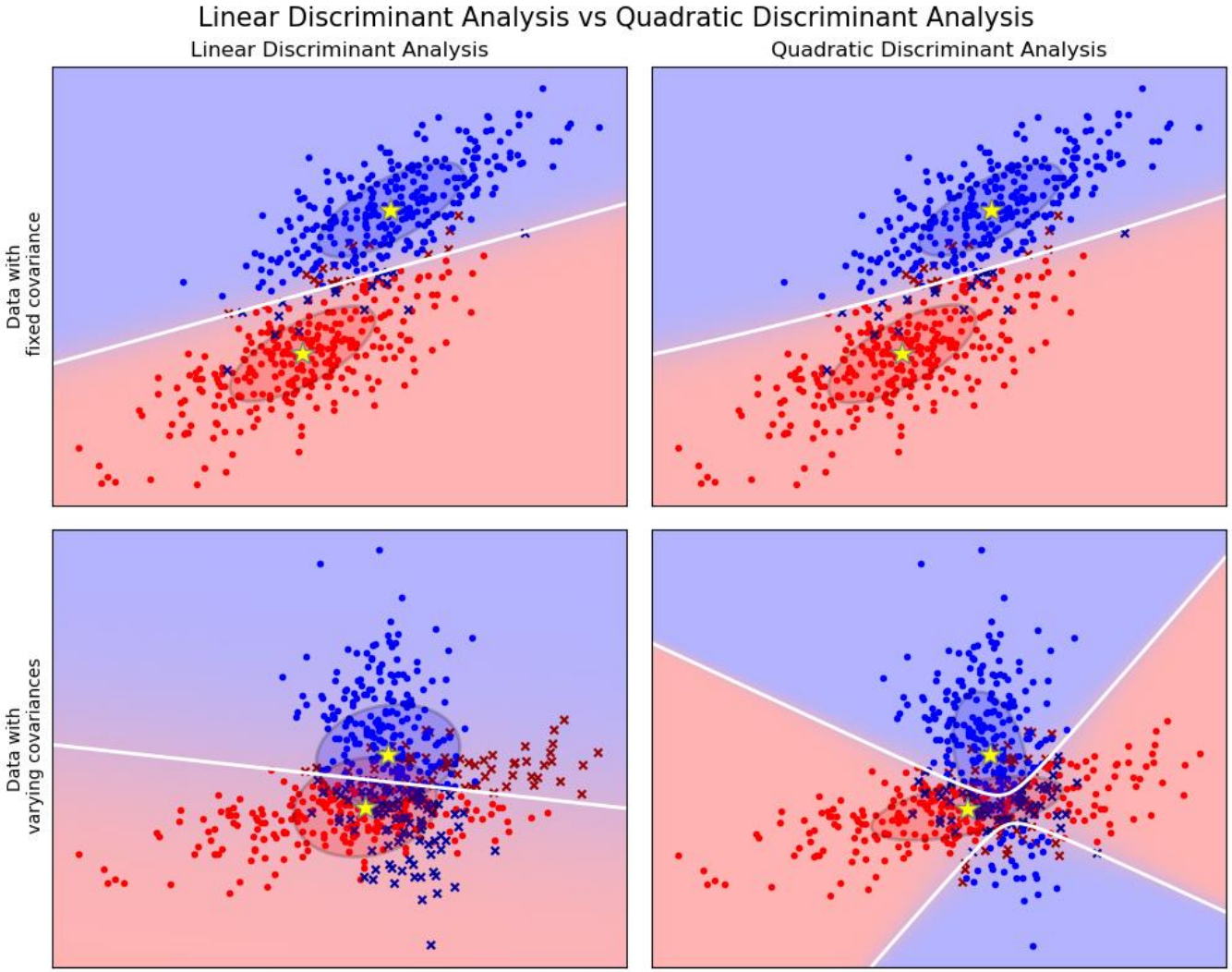
Anden eigenvektor

Første eigenvektor

Kovarians-matrix $\boldsymbol{\Sigma}$

Middelværdi $\boldsymbol{\mu}$

LDA og QDA


Linear Discriminant Analysis vs Quadratic Discriminant Analysis
Linear Discriminant Analysis
Quadratic Discriminant Analysis
Data with fixed covariance
Data with varying covariances

Naive Bayes i Scikit Learn

1.9. Naive Bayes

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. Bayes' theorem states the following relationship, given class variable y and dependent feature vector x_1 through x_n, :

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

1.9.1. Gaussian Naive Bayes ¶

`GaussianNB` implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:

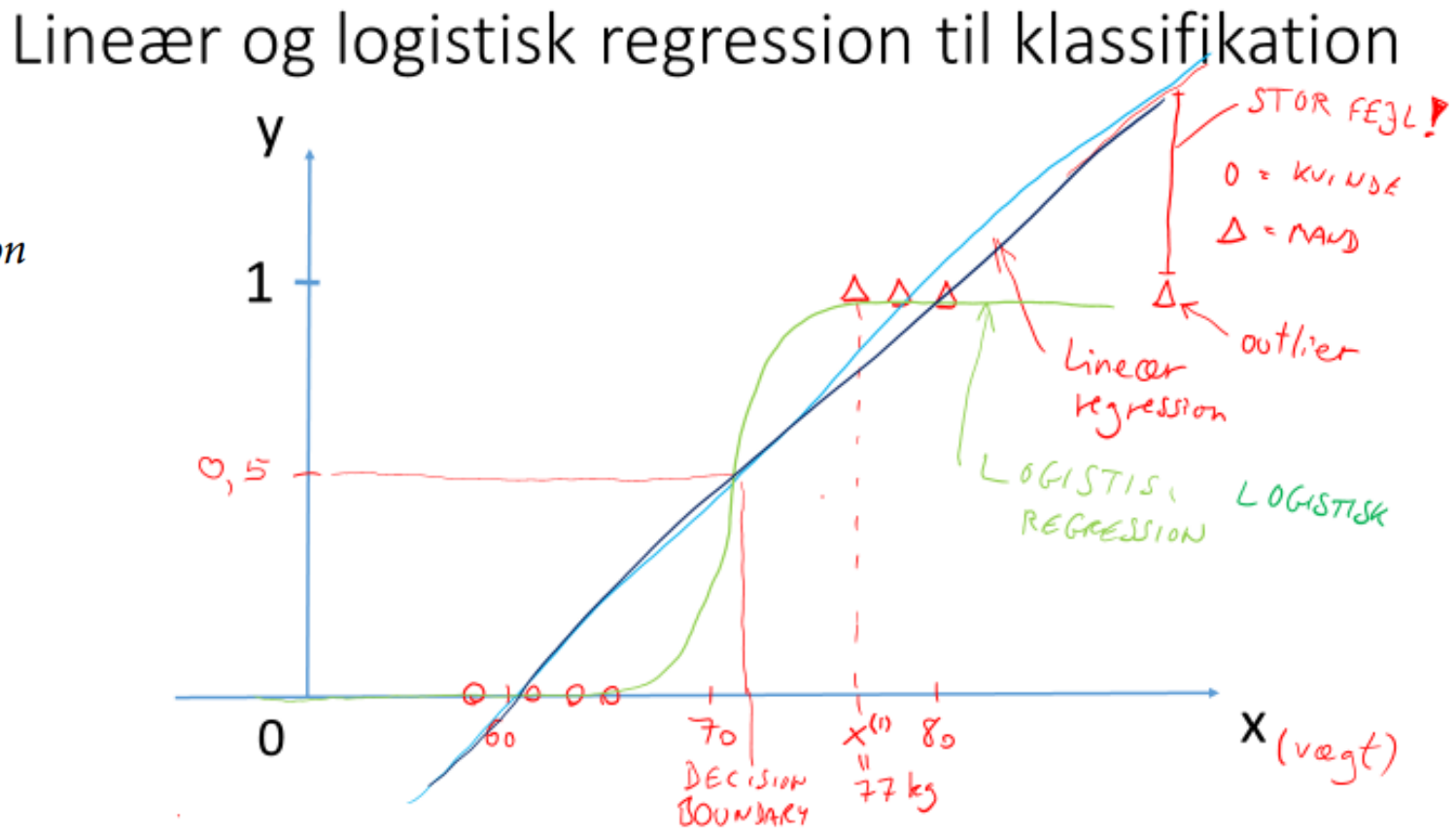
$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

The parameters σ_y and μ_y are estimated using maximum likelihood.

Logistisk regression - probabilistisk

Equation 4-14. Logistic function

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$



Maximum (log-) likelihood estimation – for normalfordeling

Log-Likelihood function

$L(\mu,\sigma|\mathbf{x}) = \ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$

μ_{ML} og σ_{ML} findes ved at maksimere $L(\mu,\sigma|\mathbf{x})$:

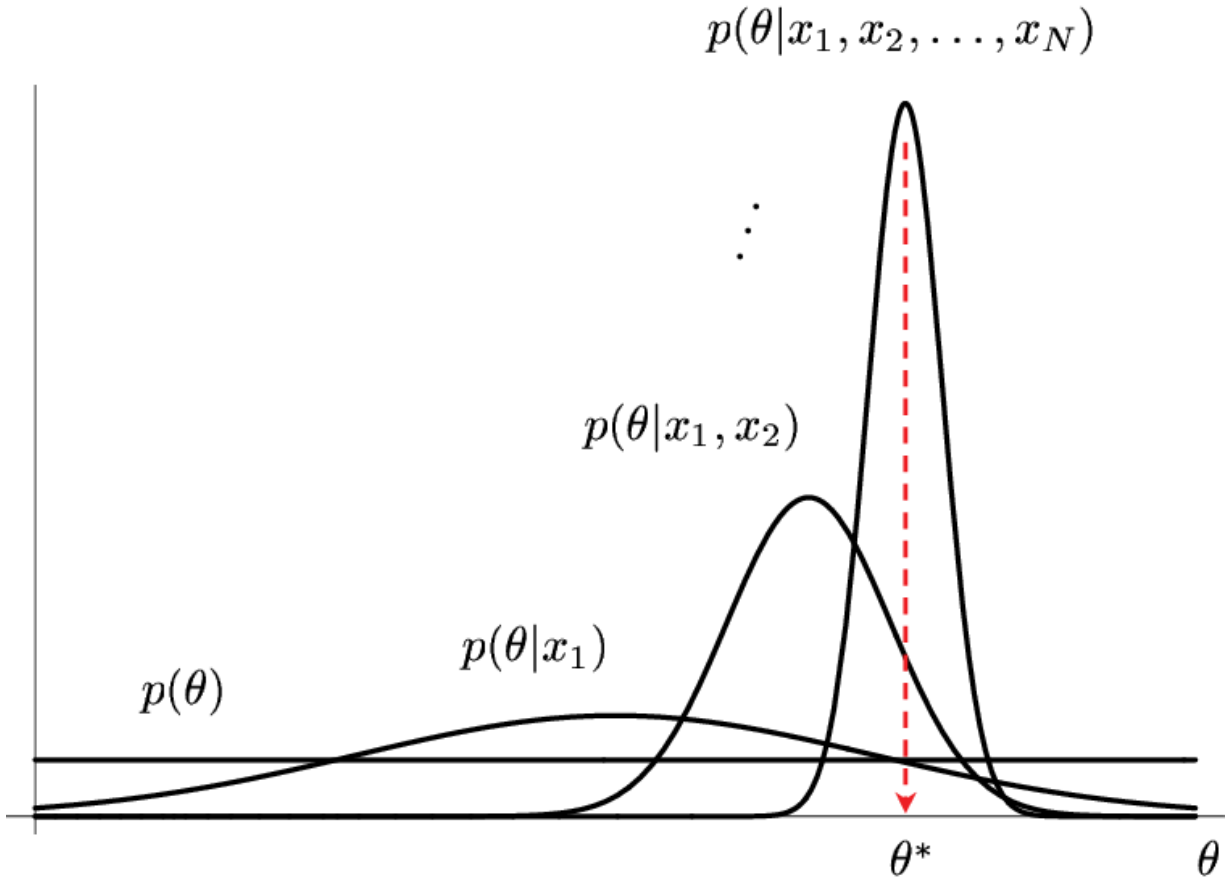
$\mu_{ML} = \arg\max_{\mu} L(\mu,\sigma|\mathbf{x})$

$\sigma_{ML} = \arg\max_{\sigma} L(\mu,\sigma|\mathbf{x})$

For normalfordeling : $\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$

Log-likelihood function = cost function !

Bayesian learning



Bayesian Learning

- Parametre (fx. μ) i pdf er også stokastiske variable med egen fordeling !
- Parametre har en *prior* fordeling
- Nogle metoder i scikit-learn har en "Bayesian" version