Bayesian Learning

Complicated derivation of known things.

- **Maximum a posteriori probability hypothesis** (MAP) (nejpravděpodobnější hypotéza)
- **Maximum likelihood hypothesis** (ML) (maximálně věrohodná hypotéza)
- **Bayesian optimal prediction** (Bayes Rate)
- Bayesian methods, bayesian smoothing
- **EM algorithm**
- **Naive Bayes model (classifier)**.

Candy Example (Russel, Norvig: Artif. Intell. a MA)

- Our favorite candy comes in two flavors: cherry and lime, both in the same wrapper.
- They are in a bag in one of following rations of cherry candies and prior probability of bags:

hypothesis (bag type)	h_1	h_2	h_3	h_4	h_5
cherry	100%	75%	50%	25%	0%
prior probability h_i	10%	20%	40%	20%	10%

- The first candy is cherry.

  MAP Which of h_i is the most probable given first candy is cherry?

  Bayes estimate What is the probability next candy from the same bag is cherry?

Maximum Aposteriory Probability Hypothesis (MAP)

- We assume large bags of candies, the result of one missing candy in the bag is negligable.
- Recall Bayes formula:

$$P(h_i|B=c) = \frac{P(B=c|h_i)\cdot P(h_i)}{\sum_{j=1,\ldots,5} P(B=c|h_j)\cdot P(h_j)} = \frac{P(B=c|h_i)\cdot P(h_i)}{P(B=c)}$$

- We look for the MAP hypothesis **maximálně aposteriorně pravděpodobná**

$$argmax_i P(h_i|B=c) = argmax_i P(B=c|h_i)\cdot P(h_i).$$

- Aposteriory probabilities of hypotheses are in the following table.

Candy Example: Aposteriory Probability of Hypotheses

index	prior	cherry ratio	cherry AND h_i	aposteriory prob. h_i
i	$P(h_i)$	$P(B = c\|h_i)$	$P(B = c\|h_i) \cdot P(h_i)$	$P(h_i\|B = c)$
1	0.1	1	0.1	0.2
2	0.2	0.75	0.15	0.3
3	0.4	0.5	0.2	**0.4**
4	0.2	0.25	0.05	0.1
5	0.1	0	0	0

- Which hypothesis is most probable?

$$h_{MAP} = argmax_i P(data|h_i) \cdot P(h_i)$$

- What is the prediction of a new candy according the most probable hypothesis h_{MAP}?

Bayesian Learning, Bayesian Optimal Prediction

- **Bayesian optimal prediction** is weigthed average of predictions of all hypotheses:

$$\begin{aligned} P(N = c|data) &= \sum_{j=1,\ldots,5} P(N = c|h_j, data) \cdot P(h_j|data) \\ &= \sum_{j=1,\ldots,5} P(N = c|h_j) \cdot P(h_j|data) \end{aligned}$$

- If our model is correct, no prediction has smaller expected error then Bayesian optimal prediction.
- We always assume i.i.d. data, independently identically distributed.
- We assume the hypothesis fully describes the data behavior. Observations are mutually conditionally independent given the hypothesis. This allows the last equation above.

Candy Example: Bayesian Optimal Prediction

i	$P(h_i \mid B = c)$	$P(N = c \mid h_i)$	$P(N = c \mid h_i) \cdot P(h_i \mid B = c)$
1	0.2	1	0.2
2	0.3	0.75	0.225
3	0.4	0.5	0.2
4	0.1	0.25	0.02
5	0	0	0
$\sum$	1		0.645

Maximum Likelihood Estimate (ML)

- Usually, we do not know prior probabilities of hypotheses.
- Setting all prior probabilities equal leads to **Maximum Likelihood Estimate, maximálně věrohodný odhad**

$$h_{ML} = argmax_i P(data|h_i)$$

- Probability of data given hypothesis = likelihood of hypothesis given data.
- Find the ML estimate:

index	prior	cherry ratio	cherry AND h_i	Aposteriory prob. h_i
i	$P(h_i)$	$P(B = c\|h_i)$	$P(B = c\|h_i) \cdot P(h_i)$	$P(h_i\|B = c)$
1	0.1	1	0.1	0.2
2	0.2	0.75	0.15	0.3
3	0.4	0.5	0.2	0.4
4	0.2	0.25	0.05	0.1
5	0.1	0	0	0

- In this example, do you prefer ML estimate or MAP estimate?
- (Only few data, over-fitting, penalization is useful. AIC, BIC)

Maximum Likelihood: Continuous Parameter θ

- New producer on the market. We do not know the ratios of candies, any h_θ, kde $\theta \in \langle 0; 1 \rangle$ is possible, any prior probabilities h_θ are possible.
- We look for maximum likelihood estimate.
- For a given hypothesis h_θ, the probability of a cherry candy is θ, of a lime candy $1 - \theta$.
- Probability of a sequence of c cherry and l lime candies is:

$$P(data|h_\theta) = \theta^c \cdot (1 - \theta)^l.$$

ML Estimate of Parameter θ

- Probability of a sequence of c cherry and l lime candies is:

$$P(data|h_\theta) = \theta^c \cdot (1-\theta)^l$$

- Usual trick is to take logarithm:

$$\ell(h_\theta; data) = c \cdot \log_2 \theta + l \cdot \log_2(1-\theta)$$

- To find the maximum of ℓ (log likelihood of the hypothesis) with respect to θ we set the derivative equal to 0:

$$\begin{aligned} \frac{\partial \ell(h_\theta; data)}{\partial \theta} &= \frac{c}{\theta} - \frac{l}{1-\theta} \\ \frac{c}{\theta} &= \frac{l}{1-\theta} \\ \theta &= \frac{c}{c+l}. \end{aligned}$$

ML Estimate of Multiple Parameters

- Producer introduced two colors of wrappers - red r and green g.
- Both flavors are wrapped in both wrappers, but with different probability of the red/green wrapper.
- We need three parameters to model this situation:

$P(B = c)$	$P(W = r \mid B = c)$	$P(W = r \mid B = l)$
θ_0	θ_1	θ_2

- Following table denotes observed frequences:

wrapper\ flavor	cherry	lime
red	r_c	r_l
green	g_c	g_l

ML Estimate of Multiple Parameters

Parameters are:

$P(B=c)$	$P(W=r\|B=c)$	$P(W=r\|B=l)$
θ_0	θ_1	θ_2

Probability of data given the hypothesis $h_{\theta_0,\theta_1,\theta_2}$ is:

$$\begin{aligned} P(data|h_{\theta_0,\theta_1,\theta_2}) &= \theta_1^{r_c}\cdot(1-\theta_1)^{g_c}\cdot\theta_0^{r_c+g_c}\cdot\theta_2^{r_l}\cdot(1-\theta_2)^{g_l}\cdot(1-\theta_0)^{r_l+g_l} \\ \ell(h_{\theta_0,\theta_1,\theta_2};data) &= r_c\log_2\theta_1 + g_c\log_2(1-\theta_1) + (r_c+g_c)\log_2\theta_0 \\ &\quad + r_l\log_2\theta_2 + g_l\log_2(1-\theta_2) + (r_l+g_l)\log_2(1-\theta_0) \end{aligned}$$

We look for maximum:

$$\begin{aligned} \frac{\partial\ell(h_{\theta_0,\theta_1,\theta_2};data)}{\partial\theta_0} &= \frac{r_c+g_c}{\theta_0} - \frac{r_l+g_l}{1-\theta_0} \\ \theta_0 &= \frac{(r_c+g_c)}{r_c+g_c+r_l+g_l} \\ \frac{\partial\ell(h_{\theta_0,\theta_1,\theta_2};data)}{\partial\theta_2} &= \frac{r_l}{\theta_2} - \frac{g_l}{1-\theta_2} \\ \theta_2 &= \frac{r_l}{r_l+g_l}. \end{aligned}$$

- Maximum Likelihood estimate is the ratio of frequences.

ML Estimate of Gaussian Distribution Parameters

- Assume x to have Gaussian distribution with unknown parameters μ a σ.
- Our hypotheses are $h_{\mu,\sigma} = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$.
- We have observed $x_1, \ldots, x_n$.
- Log likelihood is:

$$\begin{aligned} LL &= \sum_{j=1}^{N} log \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \\ &= N \cdot \left(log \frac{1}{\sqrt{2\pi}\sigma} \right) - \sum_{j=1}^{N} \frac{(x_j - \mu)^2}{2\sigma^2} \end{aligned}$$

- Find the maximum.

Linear Gaussian Distribution

- Assume random variable (feature) X.
- Assume goal variable Y with linear gaussian distribution where $\mu = b \cdot x + b_0$ and fixed variance σ^2 $p(Y|X = x) = N(b \cdot x + b_0; \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y-((b \cdot x + b_0))^2}{2\sigma^2}}$.
- Find maximum likelihood estimate of b, b_0 given a set of observations $data = \{\langle x_1, y_1 \rangle, \ldots, \langle x_N, y_N \rangle\}$.
- (Look for maximum of the logarithm of it; change the max to min with the opostite sign. Do you know this formula?)

$$argmax_{b,b_0}(log_e(\Pi_{i=1}^{N}(e^{-(y_i-(b \cdot x_i+b_0))^2}))) = argmin_{b,b_0}(?)$$

Bayesian Methods

- We specify a sampling model $P(\mathbf{Z}|\theta)$
- and a prior distribution for parameters $P(\theta)$
- then we compute

$$P(\theta|\mathbf{Z}) = \frac{P(\mathbf{Z}|\theta) \cdot P(\theta)}{\int P(\mathbf{Z}|\theta) \cdot P(\theta) d\theta},$$

  - we may draw samples
  - or summarize by the mean or mode.
  - it provides the **Bayesian optimal predictive distribution**:

$$P(z^{new}|\mathbf{Z}) = \int P(z^{new}|\theta) \cdot P(\theta|\mathbf{Z}) d\theta.$$

Example

Tossing a biased coin

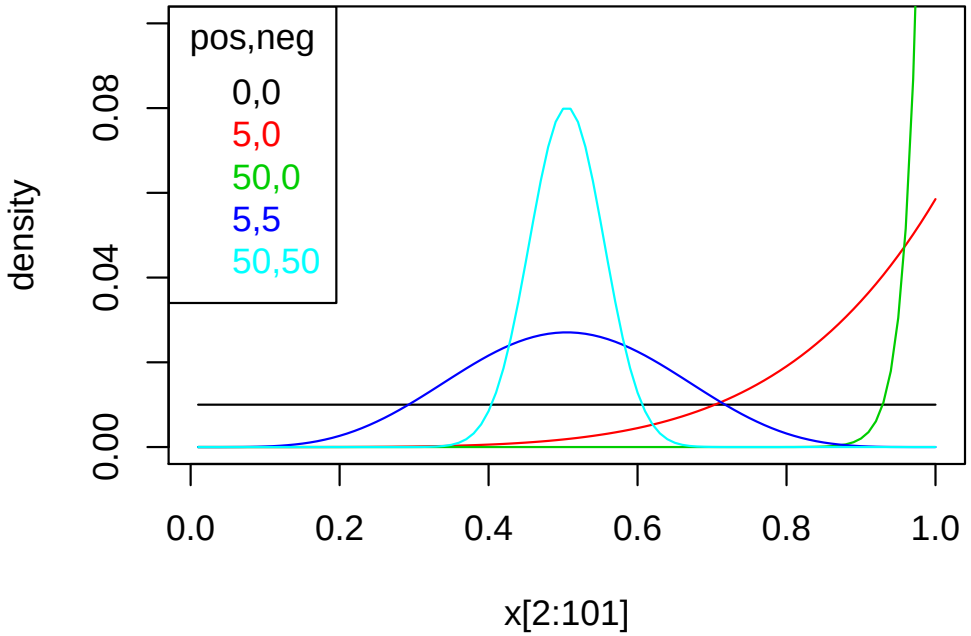
- $P(Z = head|\theta) = \theta$
- $p(\theta)$ =uniform
- $P(\theta|\mathbf{Z})$ follows the Beta distribution.

Discrete Model Parameter Learning

- For binary features, Beta function is used, $(a-1)$ is the number of positive examples, $(b-1)$ the number of negative examples.

$$beta[a,b](\theta) = \alpha\theta^{a-1}(1-\theta)^{b-1}$$

- Beta Function:



- For categorical features, Dirichlet priors and multinomial distribution is used. (Dirichlet-multinomial distribution).
- For Gaussian, μ has Gaussian prior, $\frac{1}{\sigma}$ has gamma prior (to stay in exponential family).

MAP and Penalized Methods

- MAP hypothesis maximizes:

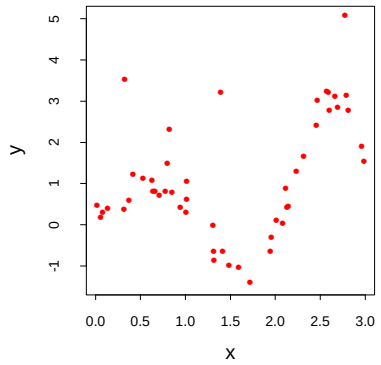
$$h_{MAP} = argmax_i P(data|h_i) \cdot P(h_i)$$

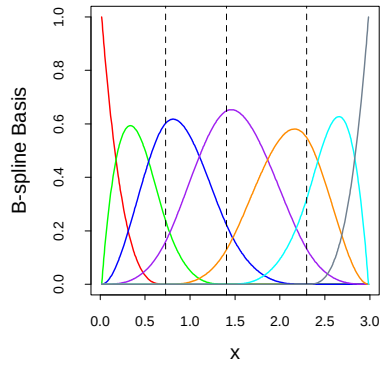
- therefore minimizes:

$$\begin{aligned} h_{MAP} &= argmax_h P(data|h)P(h) \\ &= argmin_h[-log_2 P(data|h) - log_2 P(h)] \\ &= argmin_h[-loglik + \text{complexity penalty}] \\ &= argmin_h[RSS + \text{complexity penalty}] \text{ Gaussian models} \\ &= argmax_h[loglik - \text{complexity penalty}] \text{ Categorical models} \end{aligned}$$

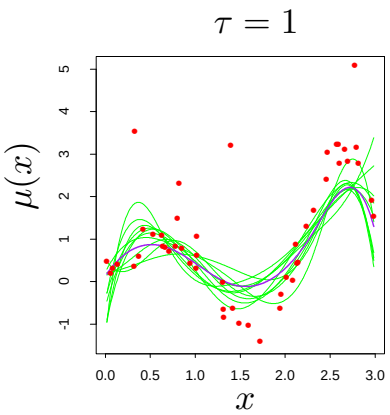
Bayesian smoothing example

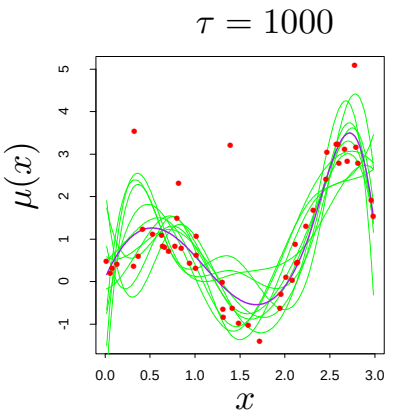
- Training data $\mathbf{Z} = \{z_i, \ldots, z_N\}$, $z_i = (x_i, y_i)$, $i = 1, \ldots, N$.
- We look for a cubic spline with three knots in quartiles of the X values. It corresponds to B-spline basis $h_j(x)$, $j = 1, \ldots, 7$.





- We estimate the conditional mean $\mathbb{E}(Y|X = x)$: $\mu(x) = \sum_{j=1}^{7} \beta_j h_j(x)$
- Let $\mathbf{H}$ be the $N \times 7$ matrix $h_j(x_i)$.
- RSS β estimate is $\hat{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$.





We assume to know σ^2, fixed x_i, we specifying prior on $\beta \sim N(0, \tau\Sigma)$.

$$\mathbb{E}(\beta|\mathbf{Z}) = (\mathbf{H}^T\mathbf{H} + \frac{\sigma^2}{\tau}\Sigma^{-1})^{-1}\mathbf{H}^T\mathbf{y}$$

$$\mathbb{E}(\mu(x)|\mathbf{Z}) = h(x)^T(\mathbf{H}^T\mathbf{H} + \frac{\sigma^2}{\tau}\Sigma^{-1})^{-1}\mathbf{H}^T\mathbf{y}.$$

Naive Bayes Model, Bayes Classifier

- Maximum Likelihood estimate is the ratio of frequences.
  - We may use smoothed estimate adding α samples to each possibility to avoid zero probabilities.
- ML estimite of a gaussian distribution parameters are the mean and the variance (or covariance matrix for multivariate distribution).
- **Naive Bayes Model, Bayes Classifier** assumes independent features given the class variable.
  - Calculate prior probability of classes $P(c_i)$
  - For each feature f, calculate for each class the probability of this feature $P(f|c_i)$
  - For a new observation of features f predict the most probable class $argmax_{c_i} P(f|c_i) \cdot P(c_i)$.

  Bayes factor
  - We can start with a comparison ratio of two classes $\frac{P(c_i)}{P(c_j)}$
  - after each observation x_p multiply it by the **bayes factor** $\frac{P(x_p|c_i)}{P(x_p|c_j)}$
  - that is:

$$\frac{P(c_i|x_1,\ldots,x_p)}{P(c_j|x_1,\ldots,x_p)} = \frac{P(c_i)}{P(c_j)} \cdot \frac{P(x_1|c_i)}{P(x_1|c_j)} \cdot \ldots \cdot \frac{P(x_p|c_i)}{P(x_p|c_j)}.$$

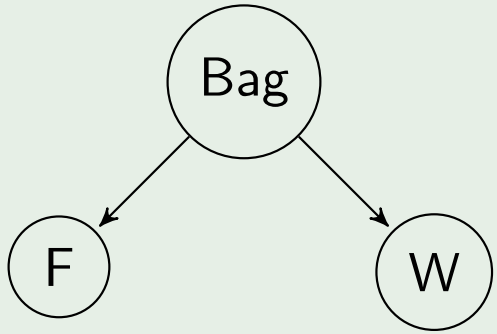
- Bayesian Networks learn more complex (in)dependencies between features.

Expectation Maximization Algorithm (EM Algorithm)

- EM algorithm estimates the maximum likelihood model based on the data with missing values.
- used in HMM
- used in clustering (Gaussian mixture model estimation)
- but not restricted to this applications
- It is a general approach to fill missing values based on the maximum likely model.

Example (EM Algorithm for Missing Data)

- Two bags of bonbons mixed together. Each bonbon has a *Wrapper* and flavor *Flavor* and may have *Holes*. Each bag had another ratio of *Wrapper* color and *Flavor*.



Bag	F	W
?	c	r
1	l	r
1	c	?
1	c	g
?	l	?

- Initialize all parameters randomly close to uniform distribution, $\theta_* \approx 0.5$.

$\mathbb{E}$ step

$w = \hat{P}(\mathbf{Z}^m \mid \theta, \mathbf{Z})$	Bag	F	W
$P_\theta(Bag = 1 \mid F = c, W = r)$	1	c	r
$P_\theta(Bag = 2 \mid F = c, W = r)$	2	c	r
1	1	l	r
$P_\theta(W = r \mid Bag = 1, F = c)$	1	c	r
$P_\theta(W = g \mid Bag = 1, F = c)$	1	c	g
1	1	c	g
$P_\theta(Bag = 1, W = r \mid F = l)$	1	l	r
$P_\theta(Bag = 1, W = g \mid F = l)$	1	l	g
$P_\theta(Bag = 0, W = r \mid F = l)$	2	l	r
$P_\theta(Bag = 0, W = g \mid F = l)$	2	l	g

M step – update θs

$$\theta_{Bag=1} \leftarrow \frac{\sum_{Bag=1} w}{\sum w}$$

$$\theta_{F=c|Bag=1} \leftarrow \frac{\sum_{Bag=1,F=c} w}{\sum_{Bag=1} w}$$

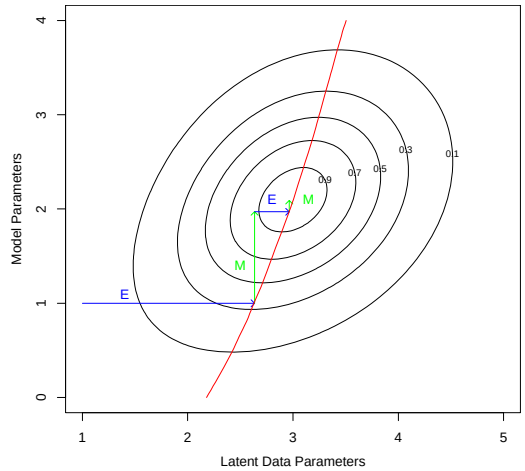
$$\theta_{F=c|Bag=2} \leftarrow \frac{\sum_{Bag=2,F=c} w}{\sum_{Bag=2} w}$$

$$\theta_{W=r|Bag=1} \leftarrow \frac{\sum_{Bag=1,W=r} w}{\sum_{Bag=1} w}$$

$$\theta_{W=r|Bag=2} \leftarrow \frac{\sum_{Bag=2,W=r} w}{\sum_{Bag=2} w}$$

EM as a Maximization-Maximization Procedure

- $\mathbf{Z}$ the observed data (the usual X with missing values)
- $\ell(\theta; \mathbf{Z})$ the log–likelihood of the model θ
- $\mathbf{Z}^m$ the latent or missing data
- $T = (\mathbf{Z}, \mathbf{Z}^m)$ the complete data with the log–likelihood $\ell_0(\theta; \mathbf{T})$.
- $\hat{P}(\mathbf{Z}^m)$,$\hat{P}(\mathbf{Z}^m|\theta, \mathbf{Z})$ any distribution over the latent data $\mathbf{Z}^m$.



- Consider the function F

$$F(\theta', \hat{P}) = \mathbb{E}_{\hat{P}}[\ell_0(\theta'; \mathbf{T})] - \mathbb{E}_{\hat{P}}[log\hat{P}(\mathbf{Z}^m)]$$

- for $\hat{P} = \hat{P}(\mathbf{Z}^m|\theta', \mathbf{Z})$ is F the log–likelihood of the observed data
  - $F(\theta', \hat{P}(\mathbf{Z}^m|\theta', \mathbf{Z})) = \mathbb{E}[\ell_0(\theta'; \mathbf{T})|\theta', \mathbf{Z}] - \mathbb{E}[\ell_1(\theta'; \mathbf{Z}^m|\mathbf{Z})|\theta', \mathbf{Z}]$

The EM Algorithm in General

$$
\begin{aligned}
P(\mathbf{Z}^m|\mathbf{Z},\theta') &= \frac{P(\mathbf{Z}^m,\mathbf{Z}|\theta')}{P(\mathbf{Z}|\theta')},\\
P(\mathbf{Z}|\theta') &= \frac{P(\mathbf{Z}^m,\mathbf{Z}|\theta')}{P(\mathbf{Z}^m|\mathbf{Z},\theta')},
\end{aligned}
$$

- In the log–likelihoods

$$\ell(\theta';\mathbf{Z}) = \ell_0(\theta';\mathbf{T}) - \ell_1(\theta';\mathbf{Z}^m|\mathbf{Z})$$

- where ℓ_1 is based on the conditional density $P(\mathbf{Z}^m|\mathbf{Z})$.
- Taking the expectation w.r.t. $\mathbf{T}|\mathbf{Z}$ governed by parameter θ gives

$$
\begin{aligned}
\ell(\theta';\mathbf{Z}) &= \mathbb{E}[\ell_0(\theta';\mathbf{T})|\theta,\mathbf{Z}] - \mathbb{E}[\ell_1(\theta';\mathbf{Z}^m|\mathbf{Z})|\theta,\mathbf{Z}]\\
&\equiv Q(\theta',\theta) - R(\theta',\theta)
\end{aligned}
$$

- $R()$ is the expectation of a density with respect the same density
  - it is maximized when $\theta' = \theta$.
- Therefore:

$$
\begin{aligned}
\ell(\theta';\mathbf{Z}) - \ell(\theta;\mathbf{Z}) &= [Q(\theta',\theta) - Q(\theta,\theta)] - [R(\theta',\theta) - R(\theta,\theta)]\\
&\geq 0.
\end{aligned}
$$

The EM Algorithm

procedure THE EM ALGORITHM:(**Z** observed data, the model(θ))
  $\hat{\theta}^{(0)} \leftarrow$ an initial guess (usually close to the uniform distribution)
  repeat
    Expectation step: at the jth step, compute

$$Q(\theta', \hat{\theta}^{(j)}) = \mathbb{E}(\ell_0(\theta'; \mathbf{T})|Z, \hat{\theta}^{(j)})$$

      as a function of the dummy argument θ'.
    Maximization step: determine the new estimate $\hat{\theta}^{(j+1)}$
      as the maximizer of $Q(\theta', \hat{\theta}^{(j)})$ over θ'.
  until convergence
  return $\hat{\theta}$
end procedure

- Full maximization is not necessary.
- We need to find a value $\hat{\theta}^{(j+1)}$ so that $Q(\hat{\theta}^{(j+1)}, \hat{\theta}^{(j)}) > Q(\hat{\theta}^{(j)}, \hat{\theta}^{(j)})$.
- Such prodecures are called **generalized EM algorithms (GEM)**.

BN example of EM algorithm (Russel, Norvig) - can be omitted

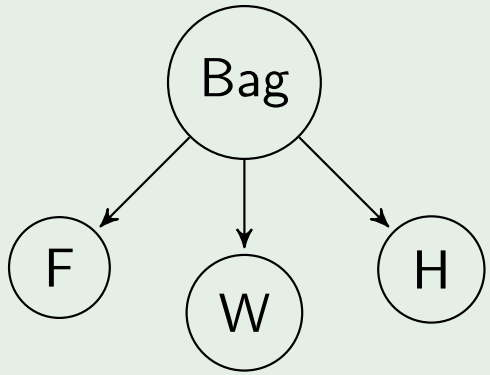
- Two bags of bonbons mixed together. Each bonbon has a *Wrapper* and flavor *Flavor* and may have *Holes*. Each bag had another ratio of *Wrapper* color, *Flavor* and *Holes*.

We can model the situation by a naive bayes model, *Bag* as the class variable.

Example

Example We have tested 1000 bonbones and observed:

	W=red		W=green	
	H=1	H=0	H=1	H=0
F=cherry	273	93	104	90
F=lime	79	100	94	167



We choose the initial parameters

$$\theta^{(0)} = 0.6,\ \theta^{(0)}_{F1} = \theta^{(0)}_{W1} = \theta^{(0)}_{H1} = 0.6,\ \theta^{(0)}_{F2} = \theta^{(0)}_{W2} = \theta^{(0)}_{H2} = 0.4$$

EM example - can be omitted

- Expectation of θ is the ratio of the expected counts

$$\theta^{(1)} = \frac{1}{N}\sum_{j=1}^{N} \frac{P(flavor_j|Bag=1)P(wrapper_j|Bag=1)P(holes_j|Bag=1)P(Bag=1)}{\sum_{i=1}^{2} P(flavor_j|Bag=i)P(wrapper_j|Bag=i)P(holes_j|Bag=i)P(Bag=i)}$$

(normalization constant **depends** on parameter values).

For the type *red*, *cherry*, *holes* we get:

$$\frac{\theta_{F1}^{(0)}\theta_{W1}^{(0)}\theta_{H1}^{(0)}\theta^{(0)}}{\theta_{F1}^{(0)}\theta_{W1}^{(0)}\theta_{H1}^{(0)}\theta^{(0)} + \theta_{F2}^{(0)}\theta_{W2}^{(0)}\theta_{H2}^{(0)}\theta^{(0)}} \approx 0.835055$$

we have 273 bonbons of this type, therefore we add $\frac{273}{N} \cdot 0.835055$.
Similarly for all seven other types and we get

$$\theta^{(1)} = 0.6124$$

EM example continued - can be omitted

- The estimate of θ_{F1} for fully observed data is $\frac{\#(Bag{=}1, Flavor{=}cherry)}{\#(Flavor{=}cherry)}$
- We have to use expected counts $Bag = 1 \& F = cherry$ and $Bag = 1$,

$$\theta_{F1}^{(1)} = \frac{\sum_{j;Flavor_j=cherry} P(Bag = 1 | Flavor_j = cherry, wrapper_j, holes_j)}{\sum_j P(Bag = 1 | cherry_j, wrapper_j, holes_j)}$$

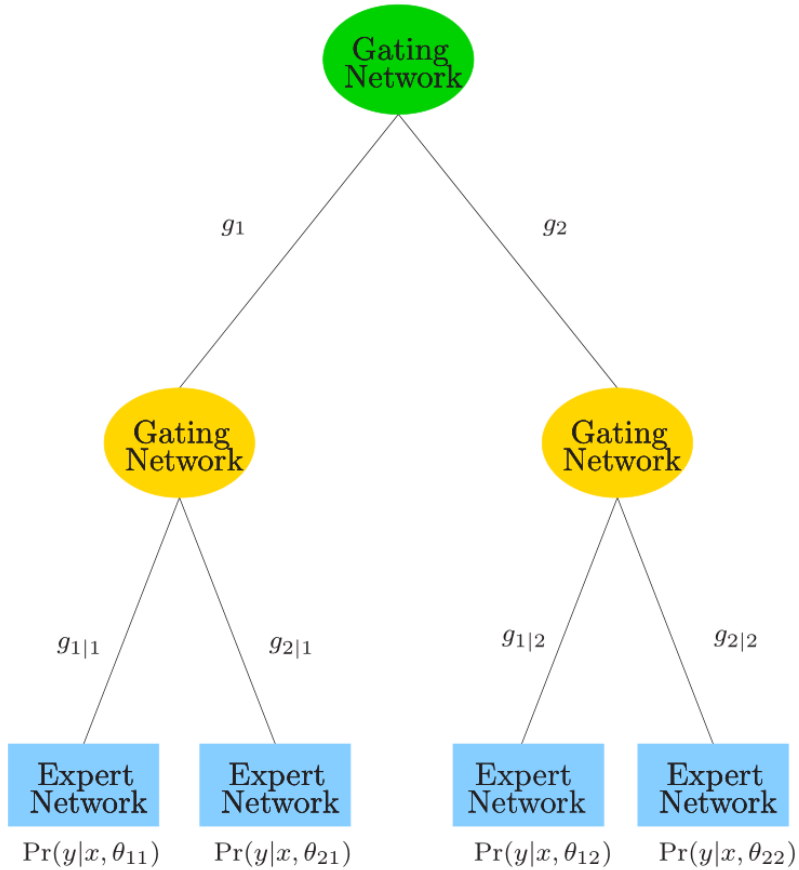
- Similarly we get:

$$\theta^{(1)} = 0.6124, \ \theta_{F1}^{(1)} = 0.6684, \theta_{W1}^{(1)} = 0.6483, \theta_{H1}^{(1)} = 0.6558,$$

$$\theta_{F2}^{(1)} = 0.3887, \theta_{W2}^{(1)} = 0.3817, \theta_{H2}^{(1)} = 0.3827.$$

Hierarchical Mixture of Experts

- a hierarchical extension of naive Bayes (latent class model)
- a decision tree with 'soft splits'
- splits are probabilistic functions of a linear combination of inputs (not a single input as in CART)
- terminal nodes called 'experts'
- non–terminal nodes are called gating network
- may be extended to multilevel.

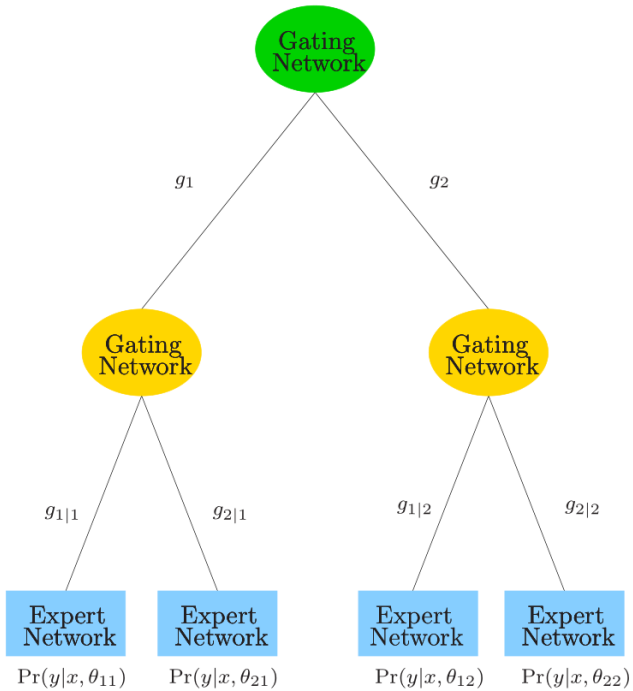


Hierarchical Mixture of Experts

- data (x_i, y_i), $i = 1, \ldots, N$, y_i continuous or categorical, first $x_i \equiv 1$ for intercepts.
- $g_i(x, \gamma_j) = \frac{e^{\gamma_j^T x}}{\sum_{k=1}^{K} e^{\gamma_k^T x}}$, $j = 1, \ldots, K$ children of the root,
- $g_{\ell|j}(x, \gamma_{j\ell}) = \frac{e^{\gamma_{j\ell}^T x}}{\sum_{k=1}^{K} e^{\gamma_{jk}^T x}}$, $\ell = 1, \ldots, K$ children of the root,
- Terminals (Experts)

Regression Gaussian linear reg. model,
$\theta_{j\ell} = (\beta_{j\ell}, \sigma^2_{j\ell})$, $Y = \beta^T_{j\ell} + \epsilon$

Classification The linear logistic reg. model:
$Pr(Y = 1|x, \theta_{j\ell}) = \frac{1}{1+e^{-\theta^T_{j\ell} x}}$



- EM algorithm
- Δ_i, $\Delta_{\ell|j}$ 0–1 latent variables – branching

E step expectations for Δ's

M step estimate parameters HME by a version of multiple logistic

Missing data (T.D. Nielsen)

Die tossed N times. Result reported via noisy telephone line. When transmission not clearly audible, record missing value:

$$4, 2, ?, 6, 5, 4, ?, 3, 4, 1, \ldots$$

"2" and "3" sound similar, therefore:

$$P(Y_i = ?|X_i = k) = P(M_i = 1|X_i = k) = \begin{cases} 1/4 & \text{k} = 2,3 \\ 1/8 & \text{k} = 1,4,5,6 \end{cases}$$

Distribution of the Y is (for fair die):

?	$\frac{1}{3}\frac{1}{4} + \frac{2}{3}\frac{1}{8} = \frac{1}{6}$
2,3	$\frac{1}{6}\frac{3}{4} = \frac{1}{8}$
1,4,5,6	$\frac{1}{6}\frac{7}{8} = \frac{7}{48}$

If we simply ignore the missing data items, we obtain as the maximum likelihood estimate for the parameters of the die:

$$\theta^* = (\frac{7}{48}, \frac{1}{8}, \frac{1}{8}, \frac{7}{48}, \frac{7}{48}, \frac{7}{48}) * \frac{6}{5} = (0.175, 0.15, 0.15, 0.175, 0.175, 0.175)$$

Incomplete data

How do we handle cases with missing values:

- Faulty sensor readings.
- Values have been intentionally removed.
- Some variables may be unobservable.

How is the data missing?
We need to take into account how the data is missing:

- **Missing completely at random** The probability that a value is missing is independent of both the observed and unobserved values (a monitoring system that is not completely stable and where some sensor values are not stored properly).
- **Missing at random** The probability that a value is missing depends only on the observed values (a database containing the results of two tests, where the second test has only performed (as a "backup test") when the result of the first test was negative).
- **Non-ignorable** Neither MAR nor MCAR (an exit poll, where an extreme right-wing party is running for parlament).

Unsupervised Learning

- No goal class (either Y nor G).
- We are interested in relations in the data:

  Clustering Are the data organized in natural clusters? (Clustering, Segmentation)
  EM algorithm for clustering
  (Dirichlet Process Mixture Models)
  (Spectral Clustering)

  Association Rules Are there some frequent combinations, implication relations? (Market Basket Analysis) *later*

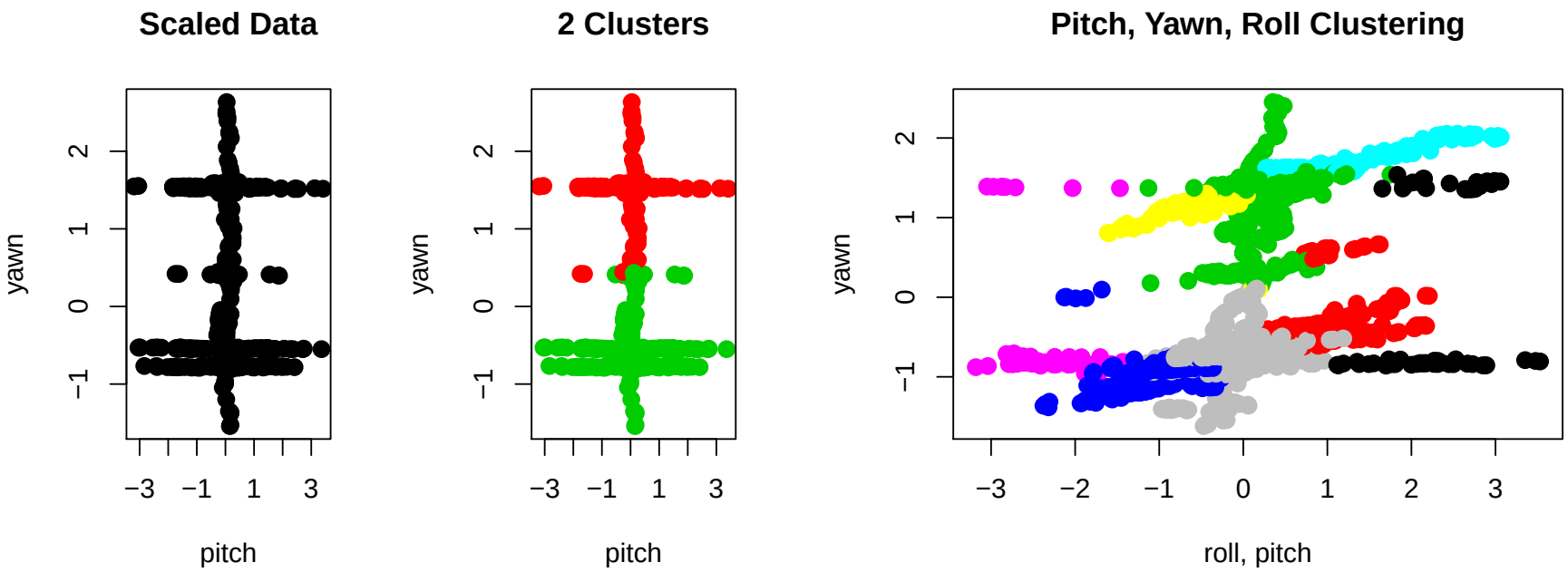
  Other The Elements of Statistical Learning Chapter 14

  SOM Self Organizing Maps

  PCA Principal Component Analysis Linear Algebra; k linear combinations of features minimizing reconstruction error (= first k principal components).
  - Principal Curves and Surfaces, Kernel and Spare Principal Components

  ICA Independent Component Analysis.

Clustering Example



- We set the color of items, no colour in train data.
- We want to assign same color to nearby points.

K – means !

K–means

1: **procedure** $K-\text{MEANS}$:(X data, K the number of clusters)
2: select randomly K centers of clusters μ_k
3: # either random data points or random points in the feature space
4: **repeat**
5: **for** each data record **do**
6: $C(x_i) \leftarrow argmin_{k \in \{1,\ldots,K\}} d(x_i, \mu_k)$
7: **end for**
8: **for** each cluster k **do** # find new centers μ_k
9: $\mu_k = \sum_{x_i : C(x_i)=k} \frac{x_i}{|C(k)|}$.
10: **end for**
11: **until** no chance in assignment
12: **end procedure**

K – means

K–means

The t iterations of K–means algorithm take $O(tkpN)$ time.

- To find global optimum is NP-hard.
- The result depends on initial values.
- May get stuck in local minimum.
- May not be robust to data sampling.
  - We may generate datasets by bootstrap method.
  - The cluster centers found in different dataset may be quite different.

  (for example, different bootstrap samples may give very different clustering results).
- Each record must belong to some cluster. Sensitive to outliers.

Distance measures

the most common distance measures:

Euclidian	$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p}(x_{ir} - x_{jr})^2}$
Hamming (Manhattan)	$d(x_i, x_j) = \sum_{r=1}^{p} \lvert x_{ir} - x_{jr} \rvert$
overlap (překrytí) categorical variables	$d(x_i, x_j) = \sum_{r=1}^{p} I(x_{ir} \neq x_{jr})$
cosine similarity	$s(x_i, x_j) = \frac{\sum_{r=1}^{p}(x_{ir} \cdot x_{jr})}{\sqrt{\sum_{r=1}^{p}(x_{jr} \cdot x_{jr}) \cdot \sum_{r=1}^{p}(x_{ir} \cdot x_{ir})}}$
cosine distance	$d(x_i, x_j) = 1 - \frac{\sum_{r=1}^{p}(x_{ir} \cdot x_{jr})}{\sqrt{\sum_{r=1}^{p}(x_{jr} \cdot x_{jr}) \cdot \sum_{r=1}^{p}(x_{ir} \cdot x_{ir})}}$

Distance – key issue, application dependent

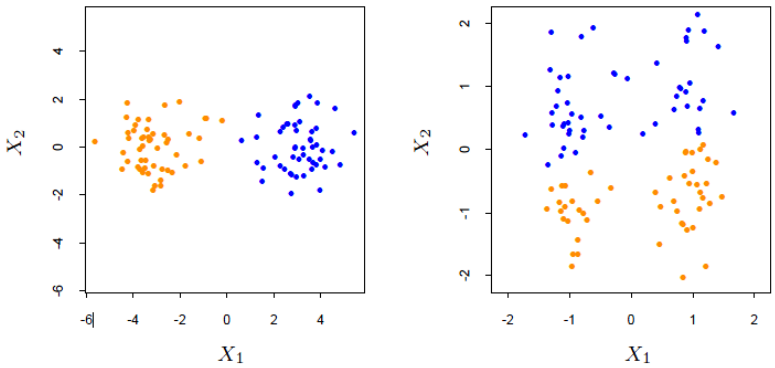
- The result depends on the choice of distance measure $d(x_i, \mu_k)$.
- The choice is application dependent.
- Scaling of the data is recommended.
- Weights for equally important attributes are: $w_j = \frac{1}{\hat{d}j}$ where

$$\hat{d}_j = \frac{1}{N^2} \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} d_j(x_{i_1}, x_{i_2}) = \frac{1}{N^2} \sum_{i_1=1}^{N} \sum_{i_2=1}^{N} (x_{i_1} - x_{i_2})^2$$

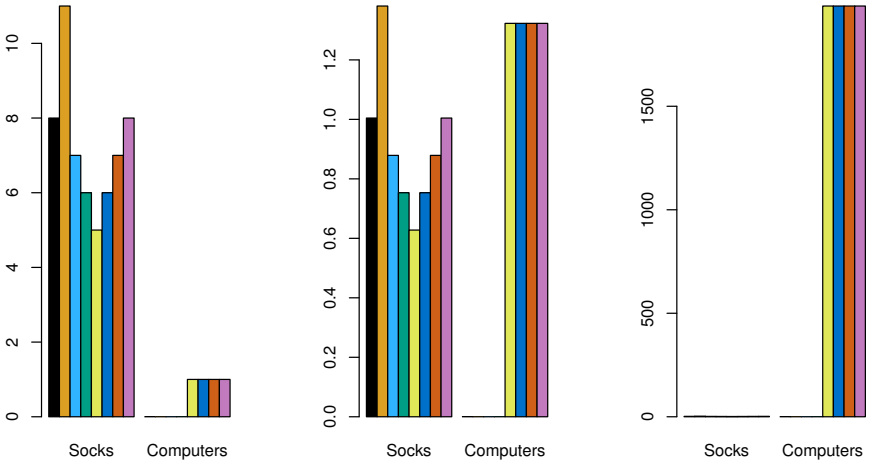
- Total distance as a weighted sum of attribute distances.
- Distance may be specified directly by a symmetric matrix, 0 at the diagonal, should fulfill triangle inequality

$$d(x_i, x_\ell) \leq d(x_i, x_r) + d(x_r, x_\ell).$$

Alternative Ideas



- Scaling may remove natural clusters
- Weighting Attributes
  - Consider internet shop offering socks and computers.
  - Compare: number of sales, standardized data, $



Number of Clusters

- We may focus on the **Within cluster variation** measure:

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k} d(x_i, x_{i'})$$

- Notice that $W(C)$ is decreasing also for uniformly distributed data.
- We look for small drop of $W(C)$ as a function of K or maximal difference between $W(C)$ on our data and on the uniform data.
- **Total** cluster variation is the sum of **between** cluster variation and **within** cluster variation

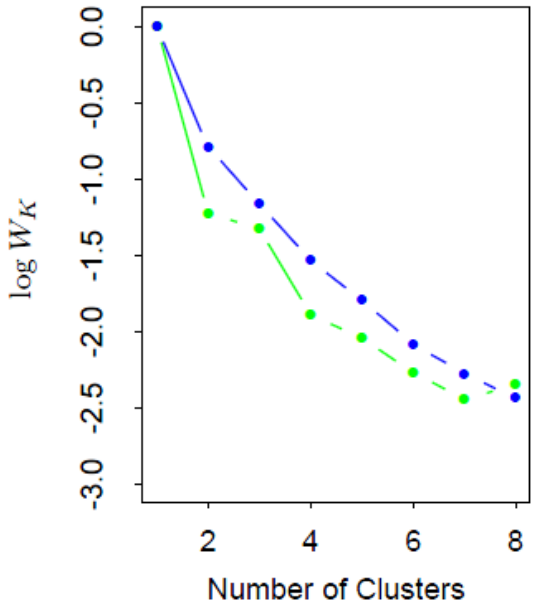
$$\begin{aligned} T(C) &= \frac{1}{2}\sum_{i,i'=1}^{N} d(x_i, x_{i'}) = W(C) + B(C) \\ &= \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\Big(\sum_{C(i')=k} d(x_i, x_{i'})\Big) + \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\Big(\sum_{C(i')\neq k} d(x_i, x_{i'})\Big) \end{aligned}$$

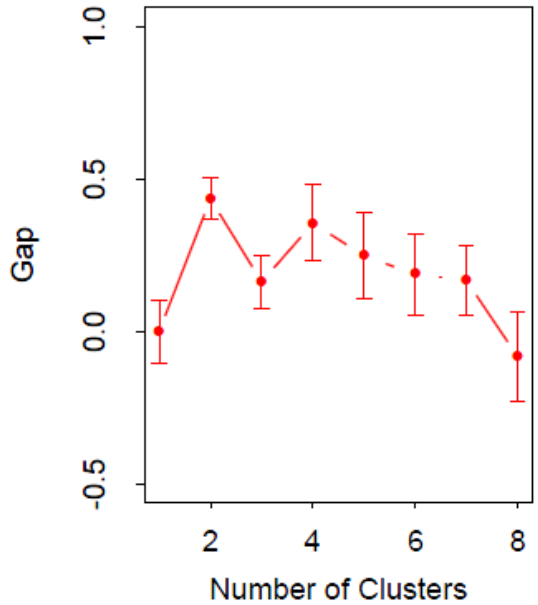
GAP function for Number of Clusters

- denote W_k the expected W for uniformly distributed data and k clusters, the average over 20 runs
- GAP is expected $log(W_k)$ minus observed $log(W(k))$

$$K^* = argmin\{k|G(k) \geq G(k+1) - s'_{k+1}\}$$

$$s'_k = s_k\sqrt{1+\frac{1}{20}} \text{ where } s_k \text{ is the standard deviation of } log(W_k)$$





Silhouette

For each data sample x_i we define

- $a(i) = \frac{1}{|C_i|-1} \sum_{j \in C_i, i \neq j} d(i,j)$ if $|C_i| > 1$
- $b(i) = min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$

Optimal number of clusters k may be selected by the SC.

Definition (Silhouette)

Silhouette s is defined

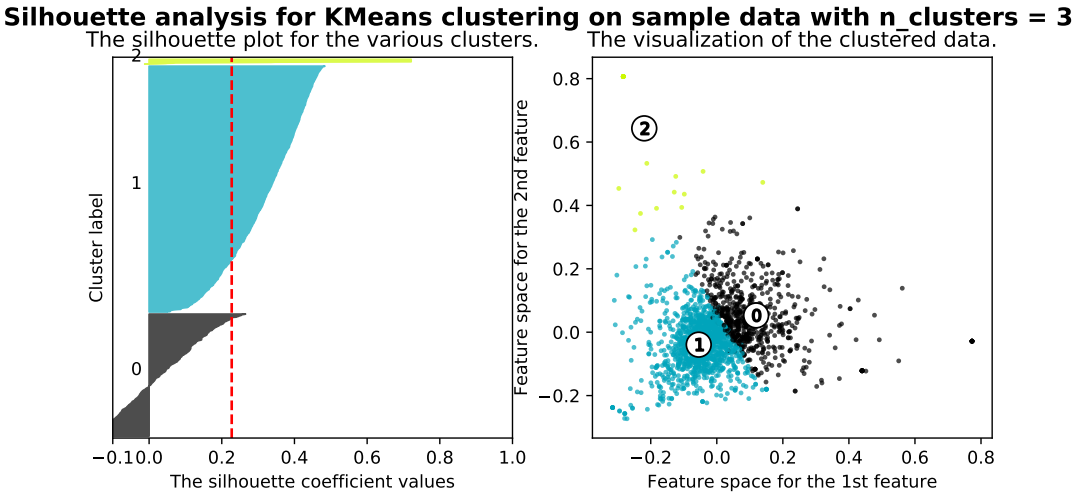
- $s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}}$ if $|C_i| > 1$
- $s(i) = 0$ for $|C_i| = 1$.

Definition (Silhouette Score)

The Silhouette score is $\frac{1}{N} \sum_i^N s(i)$.

Silhouette is always between

- $-1 \leq s(i) \leq 1$.



Note: One cluster $(-1,1),(1,1)$, other cluster $(0,-1.2),(0,-1.1)$, the point $(0,0)$ is assigned to the first cluster but has a negative silhouette. https://stackoverflow.com/a/66751204

Country Similarity Example

- Data from a political science survey: values are average pairwise dissimilarities of countries from a questionnaire given to political science students.

	BEL	BRA	CHI	CUB	EGY	FRA	IND	ISR	USA	USS	YUG
BRA	5.58										
CHI	7.00	6.50									
CUB	7.08	7.00	3.83								
EGY	4.83	5.08	8.17	5.83							
FRA	2.17	5.75	6.67	6.92	4.92						
IND	6.42	5.00	5.58	6.00	4.67	6.42					
ISR	3.42	5.50	6.42	6.42	5.00	3.92	6.17				
USA	2.50	4.92	6.25	7.33	4.50	2.25	6.33	2.75			
USS	6.08	6.67	4.25	2.67	6.00	6.17	6.17	6.92	6.17		
YUG	5.25	6.83	4.50	3.75	5.75	5.42	6.08	5.83	6.67	3.67	
ZAI	4.75	3.00	6.08	6.67	5.00	5.58	4.83	6.17	5.67	6.50	6.92

K–medoids

1: **procedure** $K-$MEDOIDS:(X data, K the number of clusters)
2: select randomly K data samples to be centroids of clusters
3: **repeat**
4: **for** each data record **do**
5: assign to the closest cluster
6: **end for**
7: **for** each cluster k **do** # find new centroids $i_k^* \in C_k$
8: $i_k^* \leftarrow argmin_{\{i:C(i)=k\}} \sum_{C(i')=k} d(x_i, x_{i'})$
9: **end for**
10: **until** no chance in assignment
11: **end procedure**

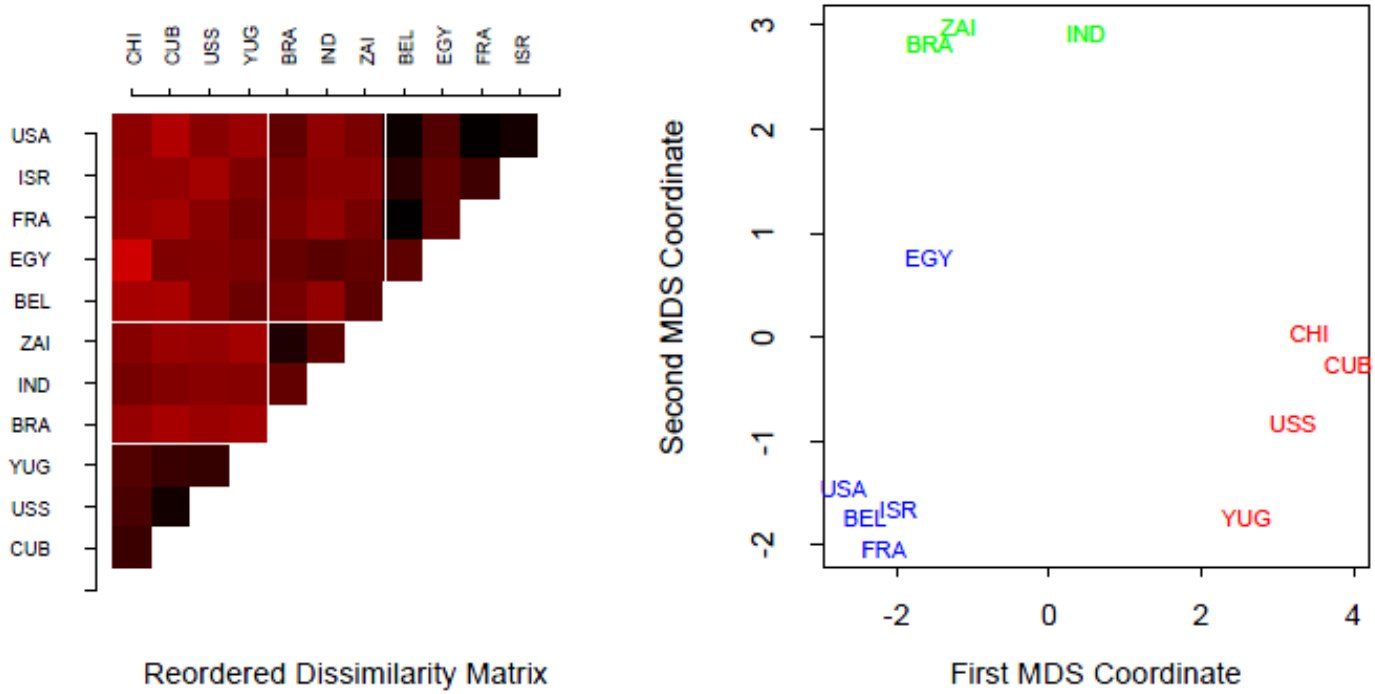
- To find a centroid requires quadratic time compared to linear k–means.
- We may use any distance, for example number of differences in binary attributes.

Complexity

The t iterations of K–medoids take $O(tkpN^2)$.

Clusters of Countries

- Survey of country dissimilarities.
- Left: dissimilarities
  - Reordered and blocked according to 3-medoid clustering.
  - Heat map is coded from most similar (dark red) to least similar (bright red).
- Right: Two-dimensional multidimensional scaling plot
  - with 3-medoid clusters indicated by different colors.



Multidimensional Scaling

- The right figure on previous slide was done by Multidimesional scaling.
- We know only distances of countries, not a metric space.
- We try to keep proximity of countries (*least squares scaling*).
- We choose the number of dimensions p.

Definition (Multidimensional Scaling)

For a given data $x_1, \ldots, x_N$ with their distance matrix d, we search $(z_1, \ldots, z_N) \in \mathbb{R}^p$ projections of data minimizing stress function

$$S_D(z_1, \ldots, z_N) = \left[\sum_{i \neq \ell} (d[x_i, x_\ell] - \|z_i - z_\ell\|)^2 \right]^{\frac{1}{2}}.$$

- It is evaluated gradiently.
- Note: Spectral clustering.

Hierarchical clustering – Bottom Up

Start with each data sample in its own cluster. Iteratively join two nearest clusters.
Measures for join

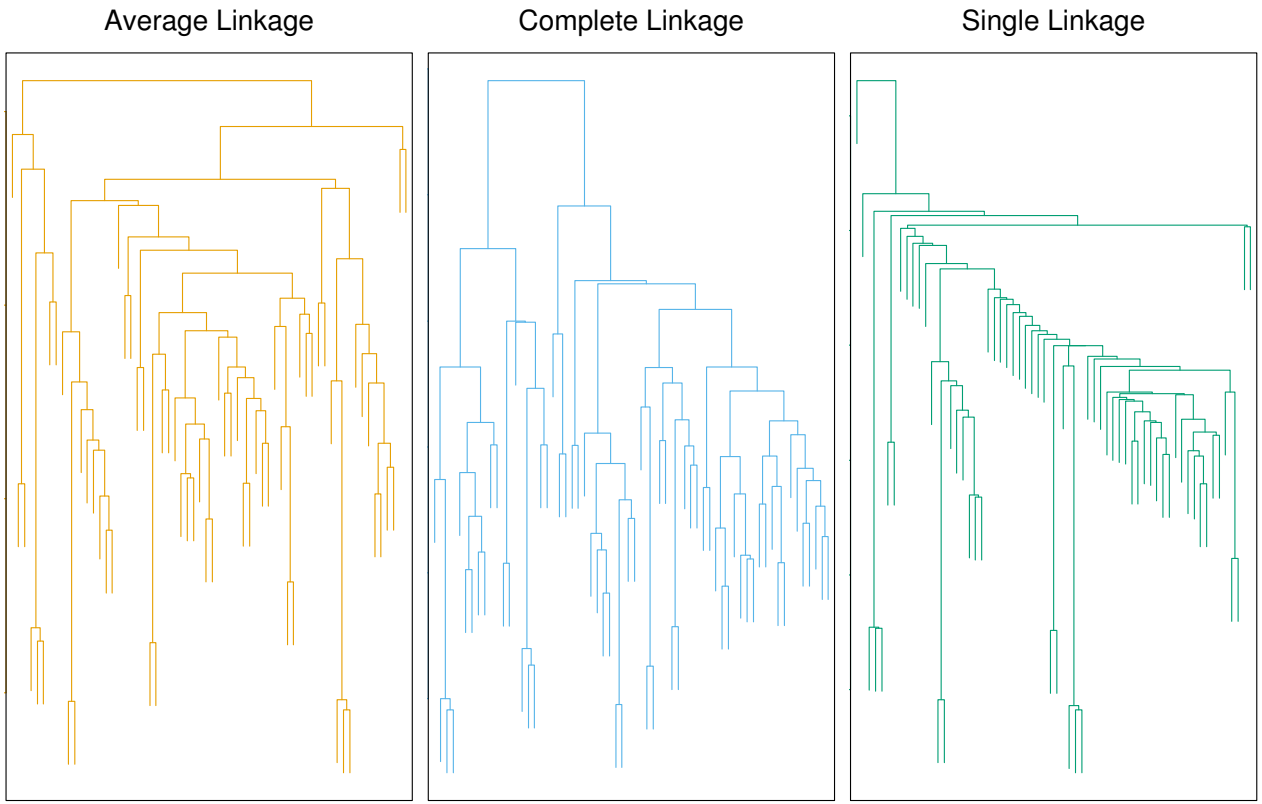
- closest points (**single linkage**)
- maximally distant points (**complete linkage**)
- **average linkage**, $d_{GA}(C_A, C_B) = \frac{1}{|C_A| \cdot |C_B|} \sum_{x_i \in C_A, x_j \in C_B} d(x_i, x_j)$
- **Ward distance** minimizes the sum of squared differences within all clusters.

$$\begin{aligned} Ward(C_A, C_B) &= \sum_{i \in C_A \cup C_B} d(x_i, \mu_{A \cup B})^2 - \sum_{i \in C_A} d(x_i, \mu_A)^2 - \sum_{i \in C_B} d(x_i, \mu_B)^2 \\ &= \frac{|C_A| \cdot |C_B|}{|C_A| + |C_B|} \cdot d(\mu_A, \mu_B)^2 \end{aligned}$$

  - where μ are the centers of clusters (A, B and joined cluster).
  - It is a variance-minimizing approach and in this sense is similar to the k-means objective function but tackled with an agglomerative hierarchical approach.

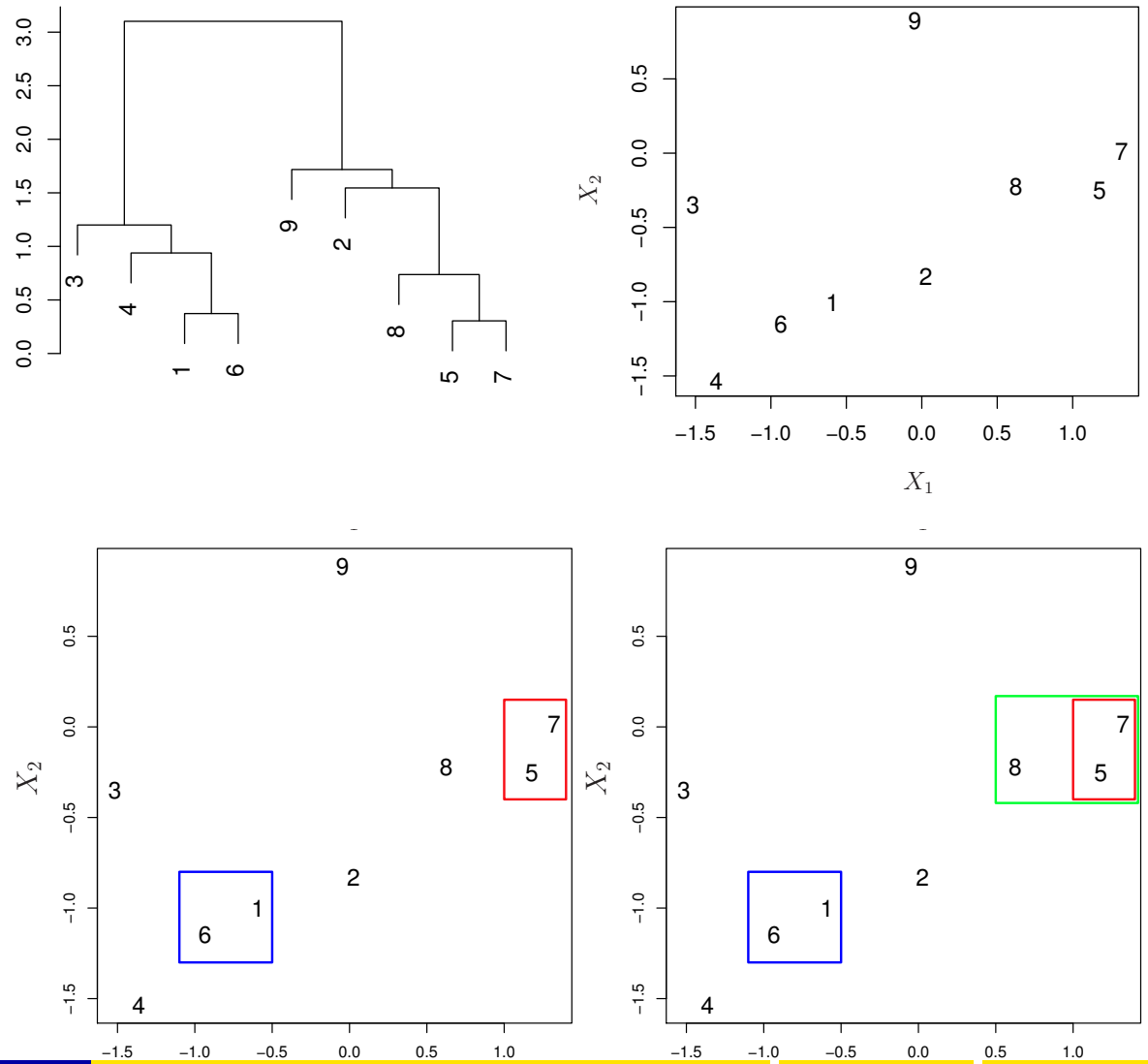
Dendrograms

- Dendrogram is the result plot of a hierarchical clustering.
- Cutting the tree of a fixed high splits samples at leaves into clusters.
  - The length of the two legs of the U-link represents the distance between the child clusters.



Interpretation of Dendrograms – 2 and 9 are NOT close

Samples fused at very bottom are close each other.



Mean Shift Clustering

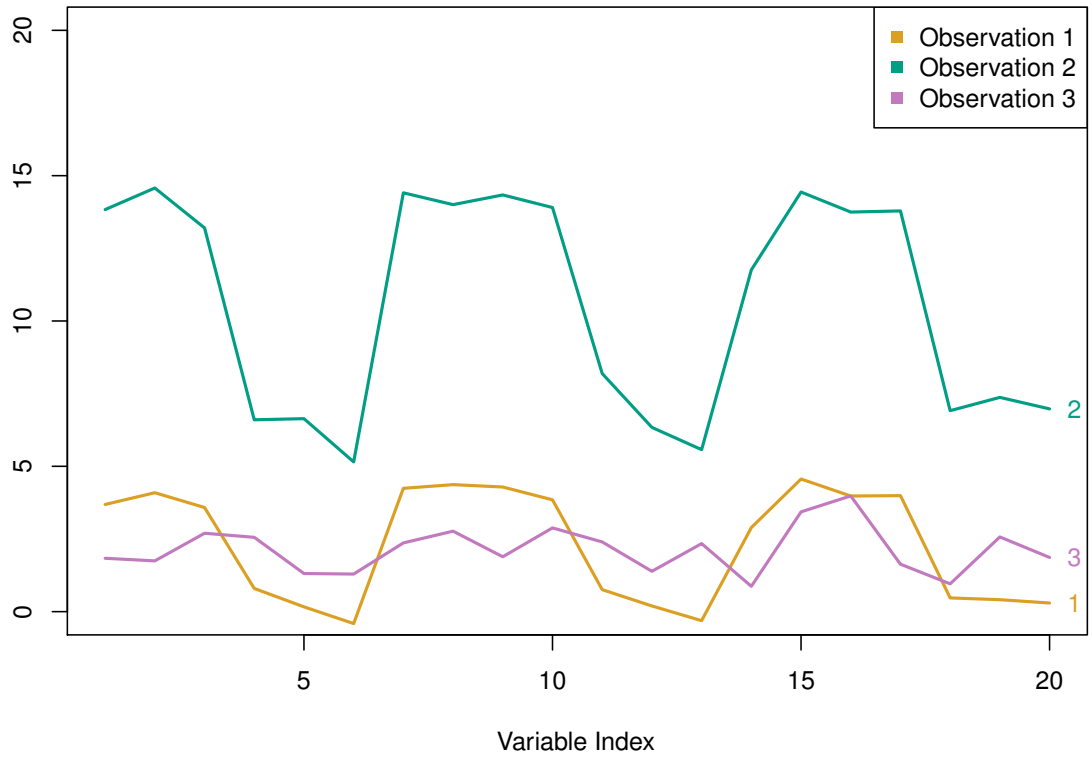
Mean Shift Clustering

1: **procedure** MEAN SHIFT CLUSTERING:(X data, $K(\cdot)$ the kernel, λ the bandwidth)
2: $\mathcal{C} \leftarrow \emptyset$
3: **for** each data record **do**
4: **repeat** # shift each mean x to the weighted average
5: $m(x) \leftarrow \frac{\sum_{i=1}^{N} K(x_i - x) x_i}{\sum_{i=1}^{N} K(x_i - x)}$
6: **until** no chance in assignment
7: add the new $m(x)$ to $\mathcal{C}$
8: **end for**
9: return prunned $\mathcal{C}$
10: **end procedure**

Kernels:

- flat kernel λ ball
- Gaussian kernel $K(x_i - x) = e^{\frac{\|x_i - x\|^2}{\lambda^2}}$

Other Distance Measures



Correlation Proximity

- Euclidian distance: Observations 1 and 3 are close.
- Correlation distance: 1 and 2 look very similar.

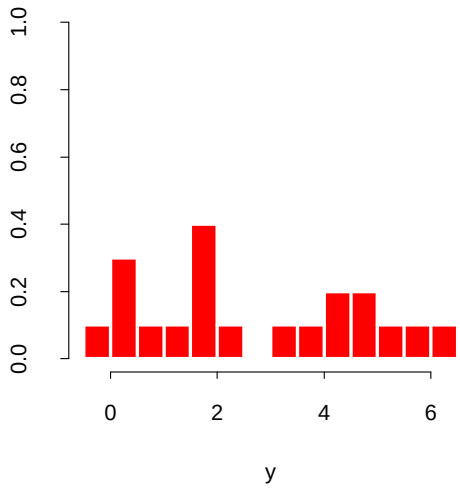
$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

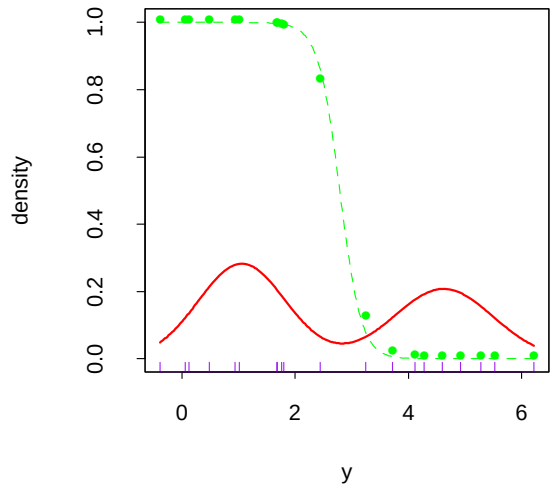
Summary

- K-means clustering - the basic one
  - the number of clusters:
  - GAP
  - Silhouette
- The distance is crucial.
  - Consider standardization or weighting the features.
- K-medoids - does need metric, just a distance
- hierarchical clustering
  - different distance measures
  - dendrogram
- other approaches (mean shift clustering, Self Organizing Maps, Spectral Clustering).

Gaussian Mixture Model

- Assume the data come from a set of k gaussian distributions
- each with
  - prior probability π_k
  - mean μ_k
  - covariance matrix Σ_k
  - $\phi_{\mu_k,\Sigma_k}(x) = \frac{1}{\sqrt{(2\pi)^p|\Sigma_k|}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}$.
- We want to find the maximum likelihood estimate of the model parameters.
- We use (more general) EM algorithm.





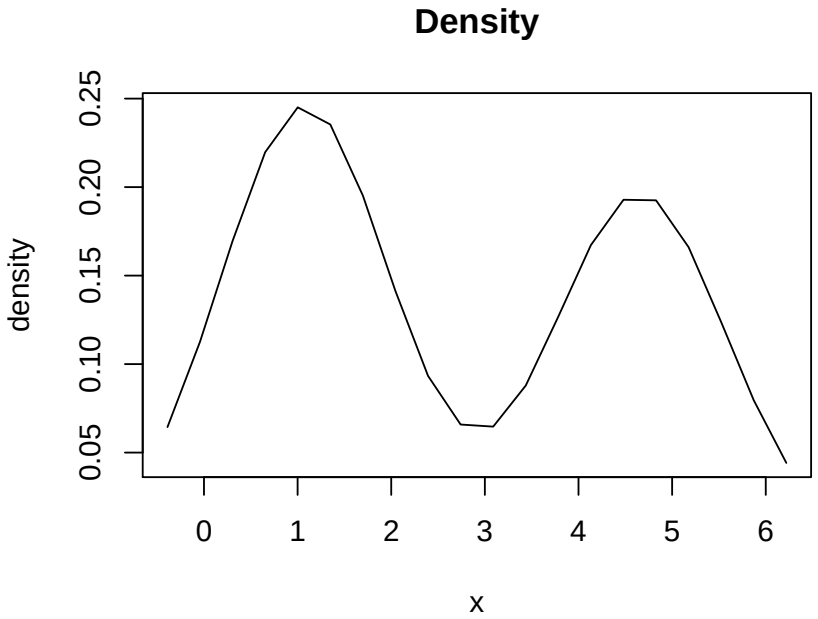
EM learning of Mixture of K Gaussians !

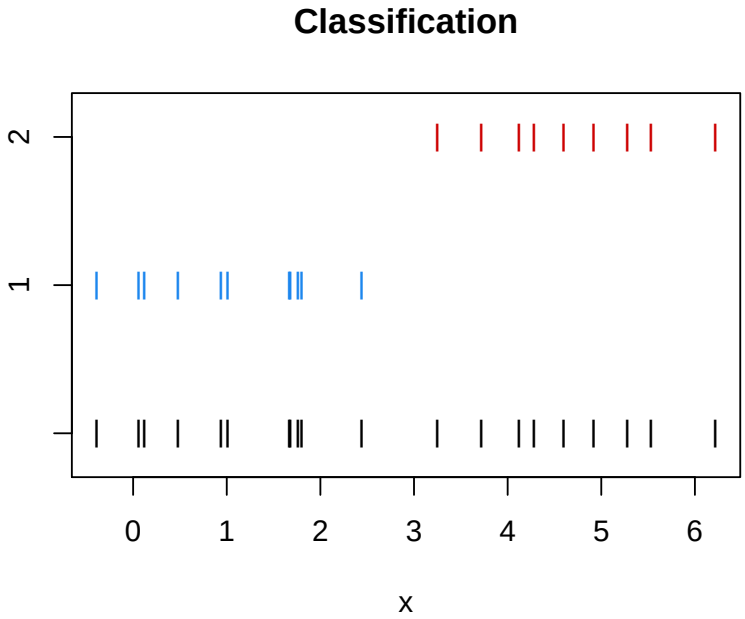
- Model parameters $\pi_1, \ldots, \pi_k, \mu_1, \ldots, \mu_k, \Sigma_1, \ldots, \Sigma_k$ such that $\sum_{k=1}^{K} \pi_k = 1$.
- **E**xpectation: weights of unobserved 'fill–ins' k of variable C:

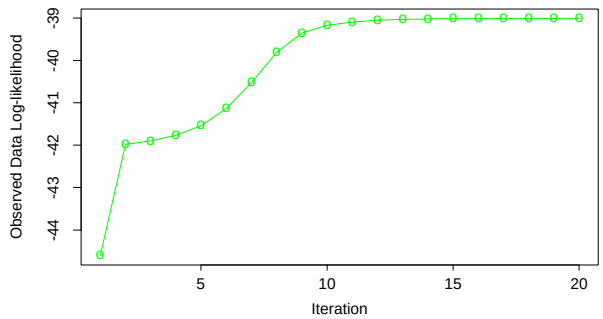
$$\begin{aligned} p_{ik} &= P(C = k|x_i) = \alpha \cdot P(x_i|C_i = k) \cdot P(C_i = k) \\ &= \frac{\pi_k \phi_{\theta_k}(x_i)}{\sum_{l=1}^{K} \pi_l \phi_{\theta_l}(x_i)} \\ p_k &= \sum_{i=1}^{N} p_{ik} \end{aligned}$$

- **M**aximize: mean, variance and cluster 'prior' for each cluster k:

$$\begin{aligned} \mu_k &\leftarrow \sum_i \frac{p_{ik}}{p_k} x_i \\ \Sigma_k &\leftarrow \sum_i \frac{p_{ik}}{p_k} (x_i - \mu_k)(x_i - \mu_k)^T \\ \pi_k &\leftarrow \frac{p_k}{\sum_{l=1}^{K} p_l}. \end{aligned}$$


Density
density
0.05
0.10
0.15
0.20
0.25
0
1
2
3
4
5
6
x


Classification
1
2
0
1
2
3
4
5
6
x


Observed Data Log-likelihood
-39
-40
-41
-42
-43
-44
5
10
15
20
Iteration

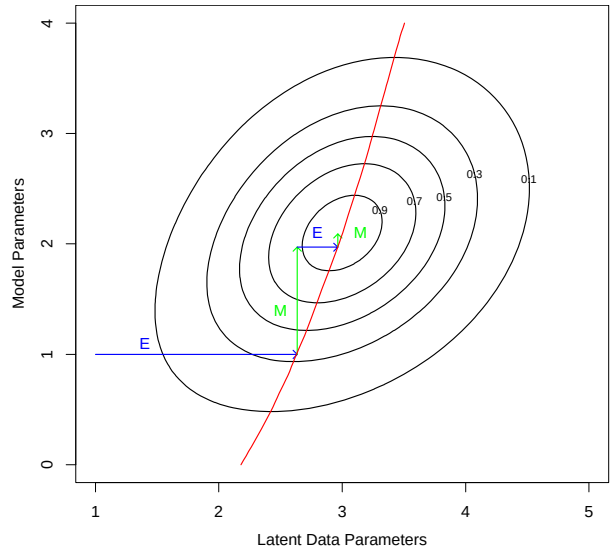

Model Parameters
0
1
2
3
4
1
2
3
4
5
Latent Data Parameters
E
M
E
M
0.9
0.7
0.5
0.3
0.1

Table of Contens