Bayesian Interpretation of Regression

$$J = \frac{1}{N} \sum_{i=1}^{N} \left(t_n - w^T \phi(x_n)\right)^2 + \lambda \cdot \sum_{j=1}^{M} w_j^2$$

$$\arg\min_w J \iff \arg\max_w \underbrace{P(x|w)}_{\text{Data likelihood}} \cdot \underbrace{P(w)}_{\text{prior}}$$

$$\| P(x|w) \sim G(t, 1)$$

$$\| P(w) \sim G(0, 1/\lambda)$$

$$\arg\max_{w} P(w|x)$$

$\equiv$ what is the value of w that

have higher probability for the data

x.

$$\arg\min_w J = \arg\max_w -J$$

$$= \arg\max_w \exp(-J)$$

$$= \arg\max_w \exp\left(\frac{1}{N}\sum_{i=1}^{N} e_i^2 + \lambda \cdot \sum_{j=1}^{M} w_j^2\right)$$

$$= \arg\max_w \underbrace{\exp\left(\frac{1}{N}\sum_{i=1}^{N} e_i^2\right)}_{\sim G(e|0,1)} \cdot \underbrace{\exp\left(\lambda \cdot \sum_{j=1}^{M} w_j^2\right)}_{\sim G(0,1/\lambda)}$$

Bayesian interpretation says we can model a given set of data $\{(x_i, t_i)\}_{i=1}^{N}$ using any distributional form for the data likelihood and for the prior distribution.

$$\arg\max_w P(x|w) \cdot P(w)$$

$$\propto \arg\max_w P(w|x) \quad \text{(from Bayes' theorem)}$$

Online setting

↳ New samples are arriving every day.

↳ If Posterior - Prior have a Conjugate Prior Relationship.

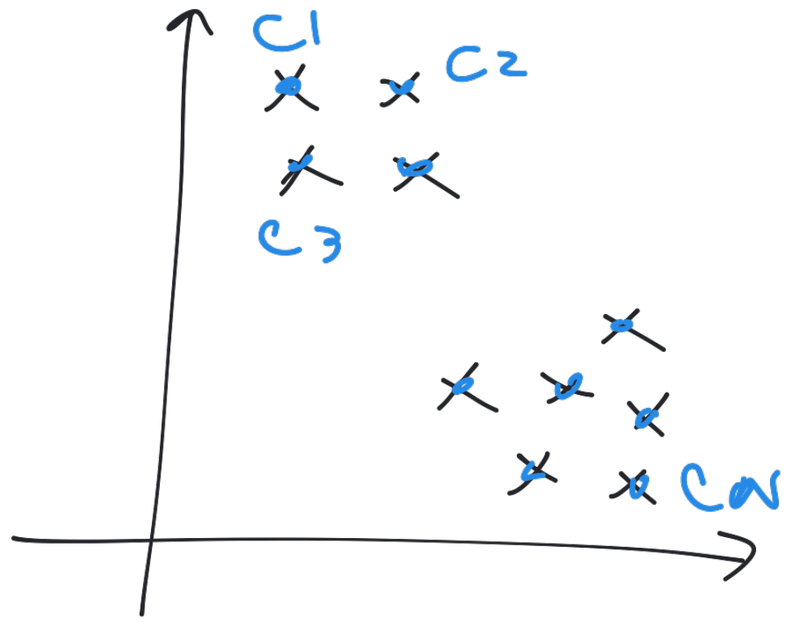
→ as new data comes in,

We can updat the prior with the posterior.

→ Posterior will represent the new data informative prior.

K-Means

$\{x_i\}_{i=1}^{N}$



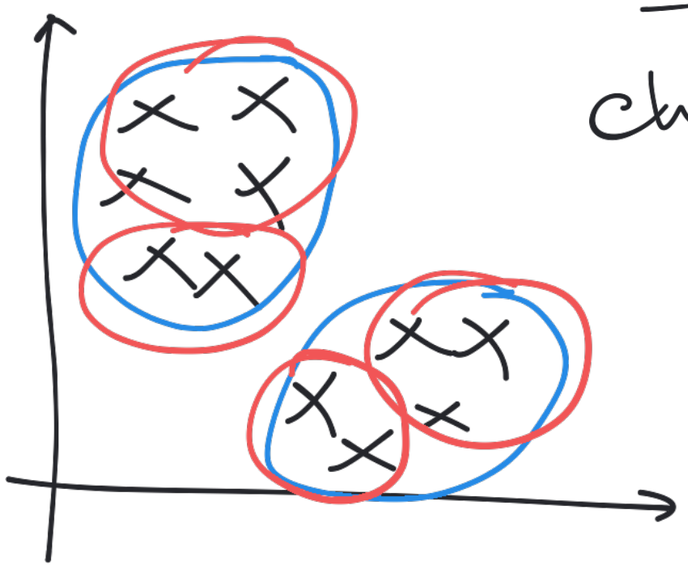
$$J = \sum_{i=1}^{N} \sum_{j=1}^{K} u_{ij} \cdot d^2(x_i, c_j) = \sum_{i=1}^{N} \sum_{j=1}^{K} u_{ij} \|x_i - c_j\|_2^2$$

$$u_{ij} \in \{0, 1\}, \quad \sum_{j=1}^{K} u_{ij} = 1$$

when $K = N$: $J = 0$!

What should we do to learn # clusters K?

↳ Internal Criteria for our cluster validity metric



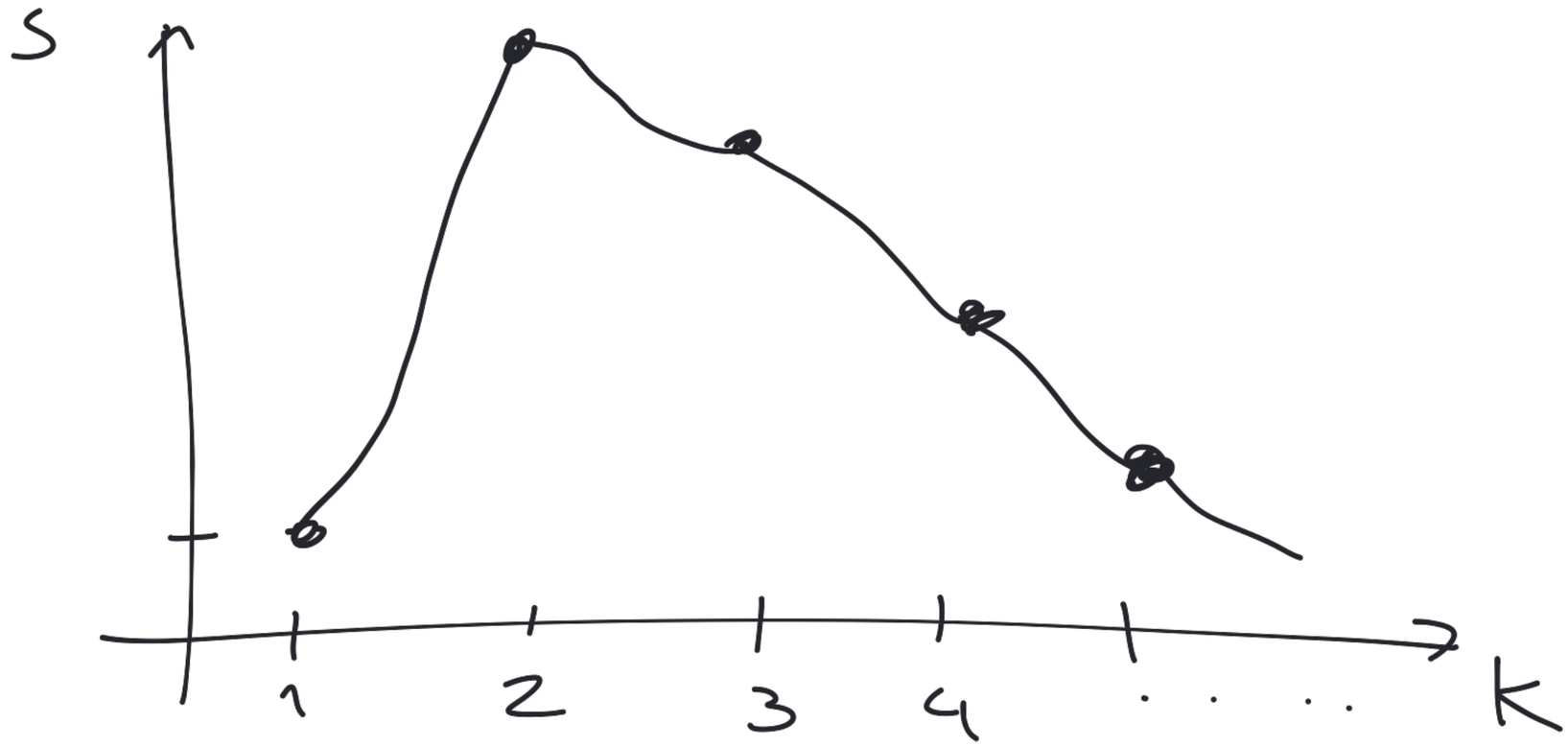
K=2 , K=4

① Compact: all points assigned to the same cluster are near each other

② Distance between clusters to be large.

Silhouette index

$-1 \leq S \leq 1$



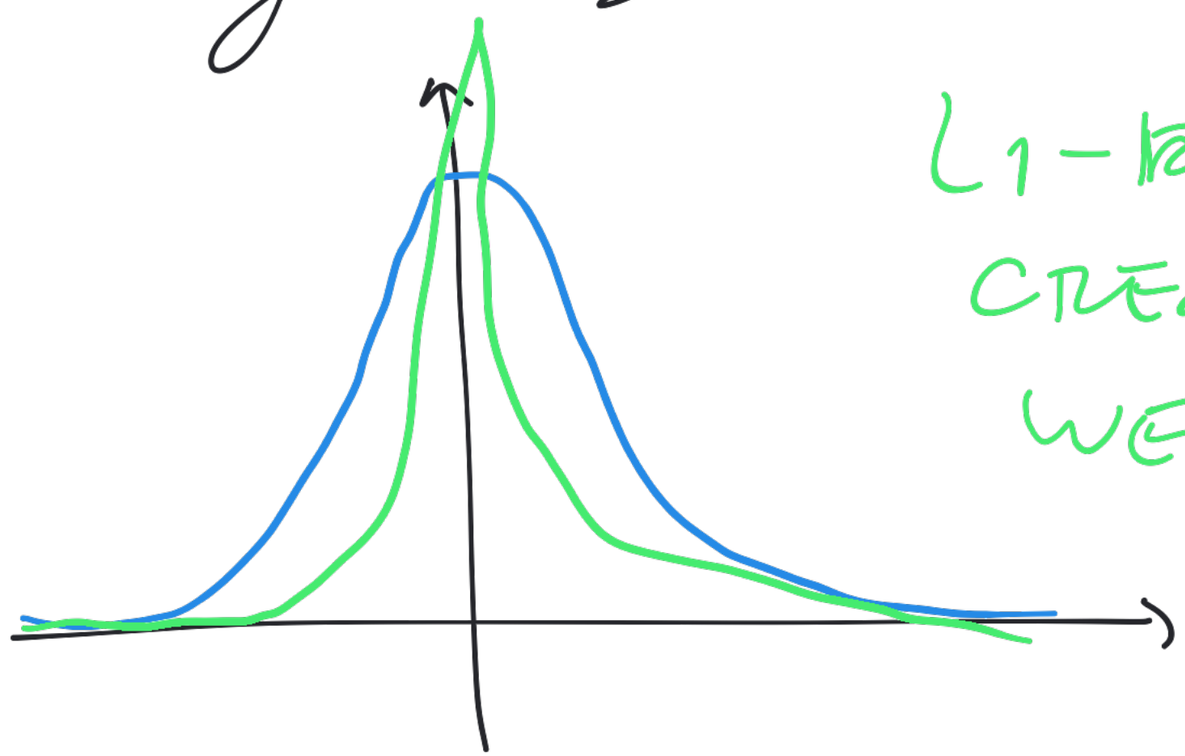
We choose K with maximum Silhouette value.

→ in this case $\underline{K = 2}$.

From Bayesian interpretation:

L2 - Regularizer is $G(0, 1/\lambda)$

L1 - Regularizer is $\mathcal{L}(0, 1/\lambda)$



L1-Reg. will create sparse weight vectors

ML

SUPERVISED ML $\{(x_i, t_i)\}_{i=1}^{N}$

- REGRESSION
- CLASSIFICATION

UNSUPERVISED ML $\{x_i\}_{i=1}^{N}$

- CLUSTERING
- DIMENSIONALITY REDUCTION

For iteration t:

Data X

prior $\sim$ Beta $(\mu \mid \alpha, \beta)$

$P(\mu)$

data likelihood $\sim B(\mu)$

$P(X \mid \mu)$

posterior = data likelihood $\times$ prior

$P(\mu \mid X)$

prior $\leftarrow$ posterior

$\alpha \leftarrow \alpha + M$

$\beta \leftarrow \beta + L$