Probabilistic parameter estimation

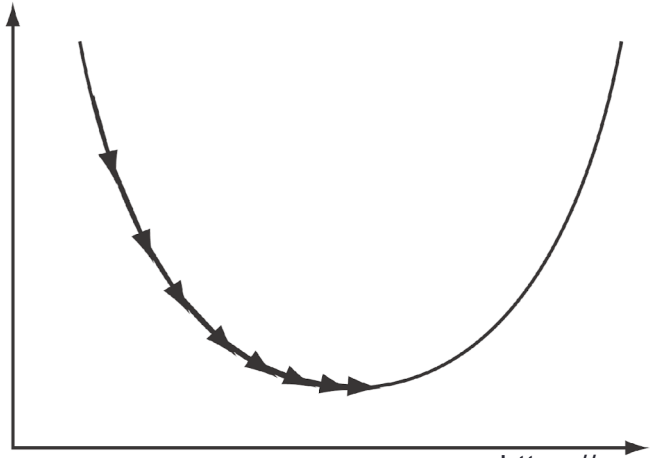
MLE, MAP Naïve Bayes
GMM and EM

Loose Ends from HW

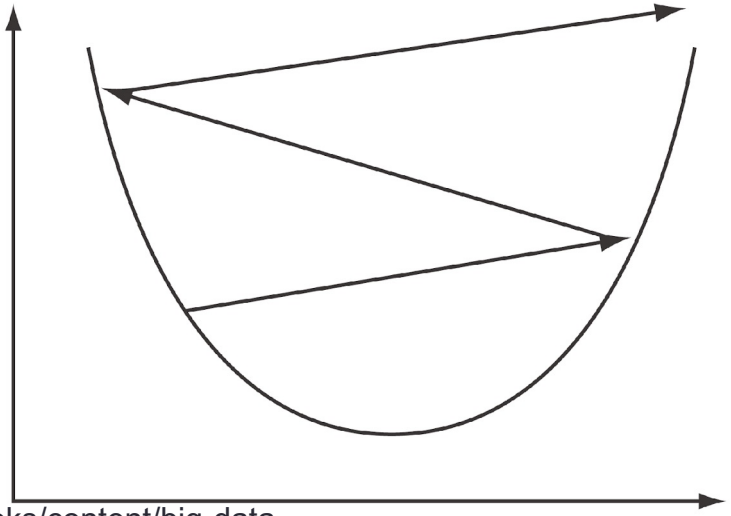
- How to select r? (The learning rate)

$$\theta_j \Leftarrow \theta_j + r\Sigma_{i=1}^{m}(y_i - \theta^T \mathbf{x}_i)x_i^{(j)}$$

Small learning rate



Large learning rate



https://www.packtpub.com/books/content/big-data

r too small and the model converges slowly

r too large and the model diverges

Learning rate issues

- Typically, r is normalized with the amount of training examples in a mini-batch. (Divide by m)

$$\theta_j \Leftarrow \theta_j + r\Sigma_{i=1}^{m}(y_i - \theta^T \mathbf{x}_i)x_i^{(j)}$$

- Typical values are 0.1-0.001
- Usually have a decay over time

Scaling the input data

- We use age, passenger class, gender, and embark as our input.
- Age has a lot more variance (0.42 – 80) than the other data.
- This makes parameter initialization hard and makes the learning rate selection hard.
- $h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$

Scaling the input data

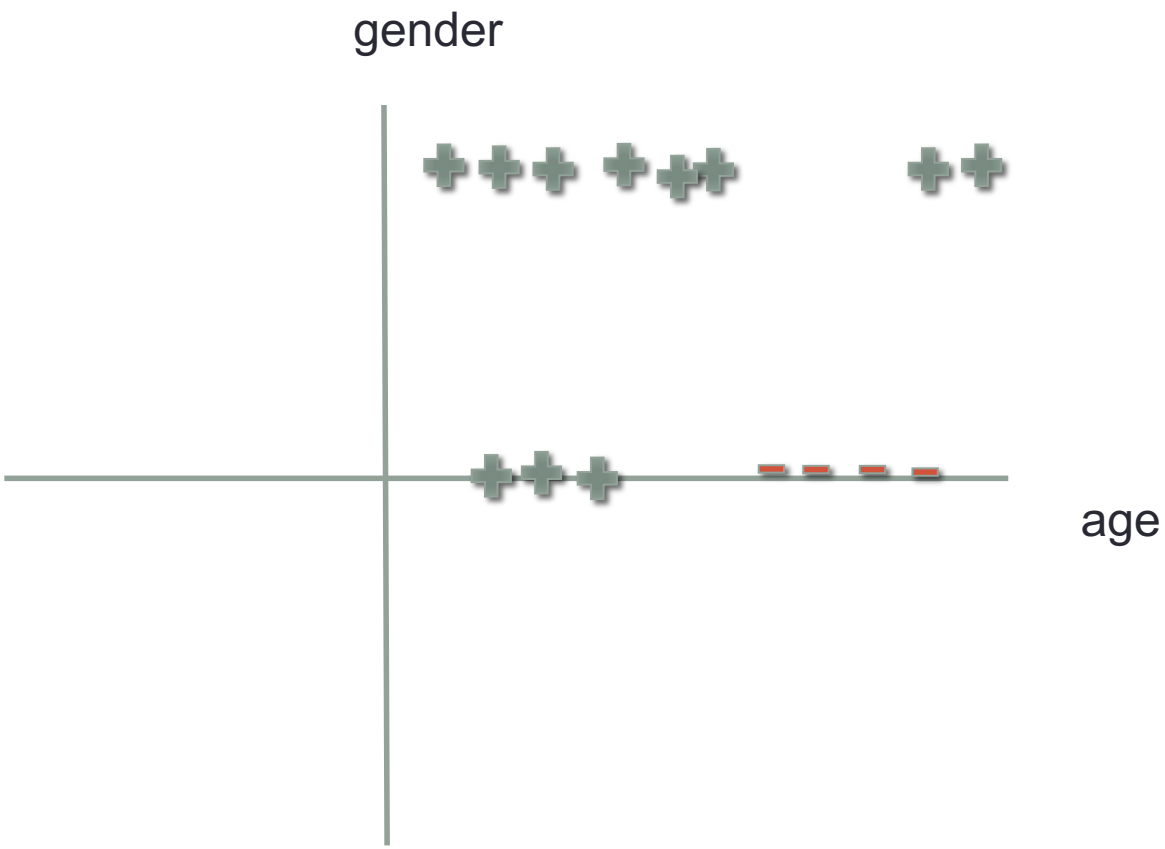
- Scale all input data to be in the same range
- Using statistics from training data
  - Scale to [-1,1]
  - Scale to [0, 1]
  - Scale to standard normal
- Don’t forget to apply the same scaling to the test data

Feature selection

- Most likely you will get better results with just two features.
- This is the importance of feature selection.
- Knowing what good features to select is not trivial
- Approaches for feature selection (or for not having to do feature selection)
  - Cross validation
  - Random forest
  - Boosting

Feature engineering

- Logistic regression is a linear classification

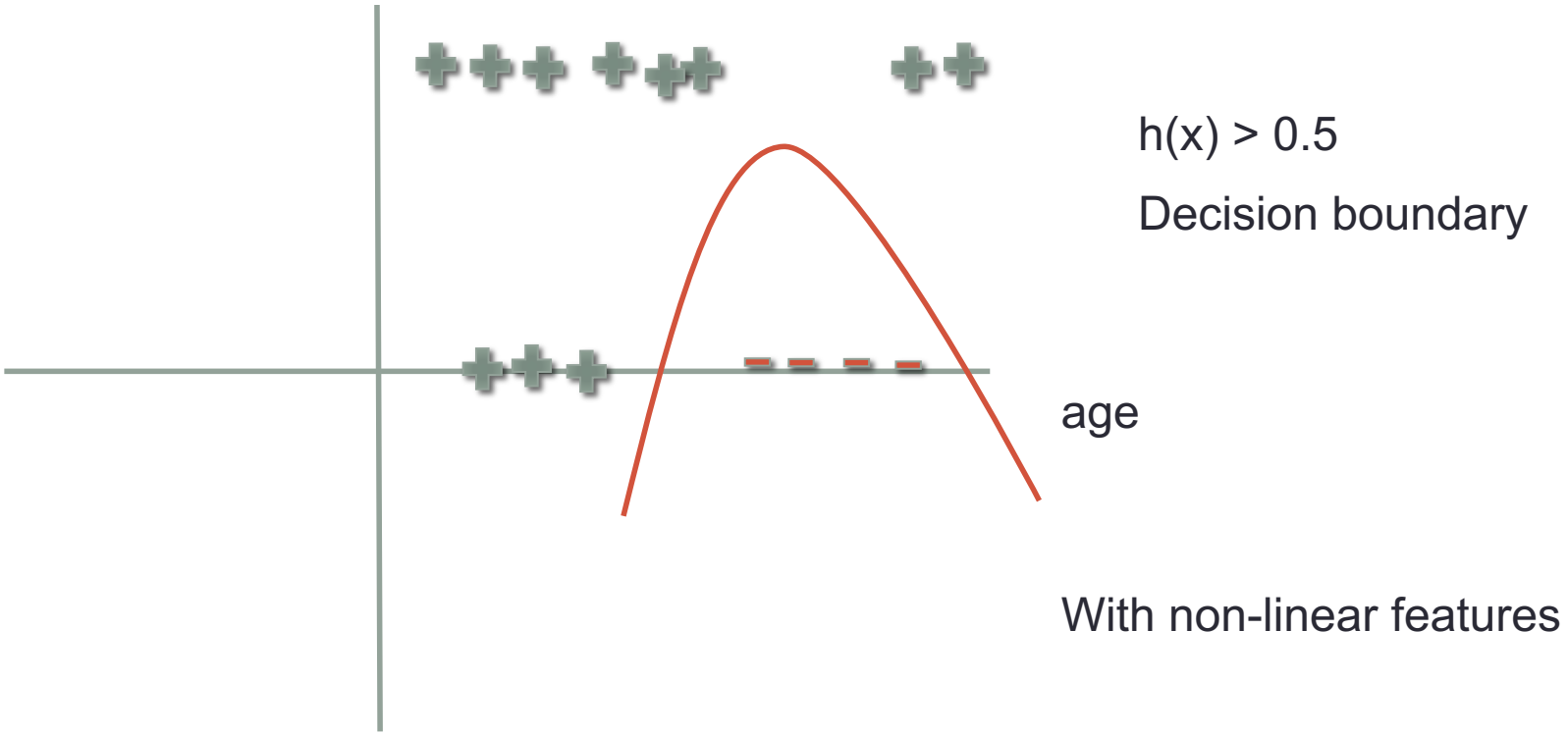


Feature engineering

- Logistic regression is a linear classification

Feature engineering

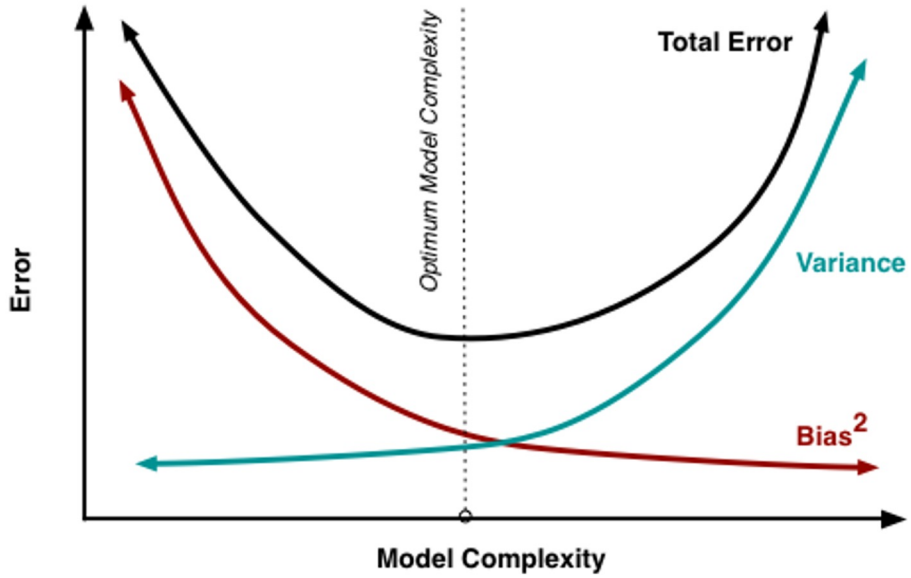
- Add non-linear features to get non-linear decision boundaries



This is also a form of feature selection (more specifically feature engineering)

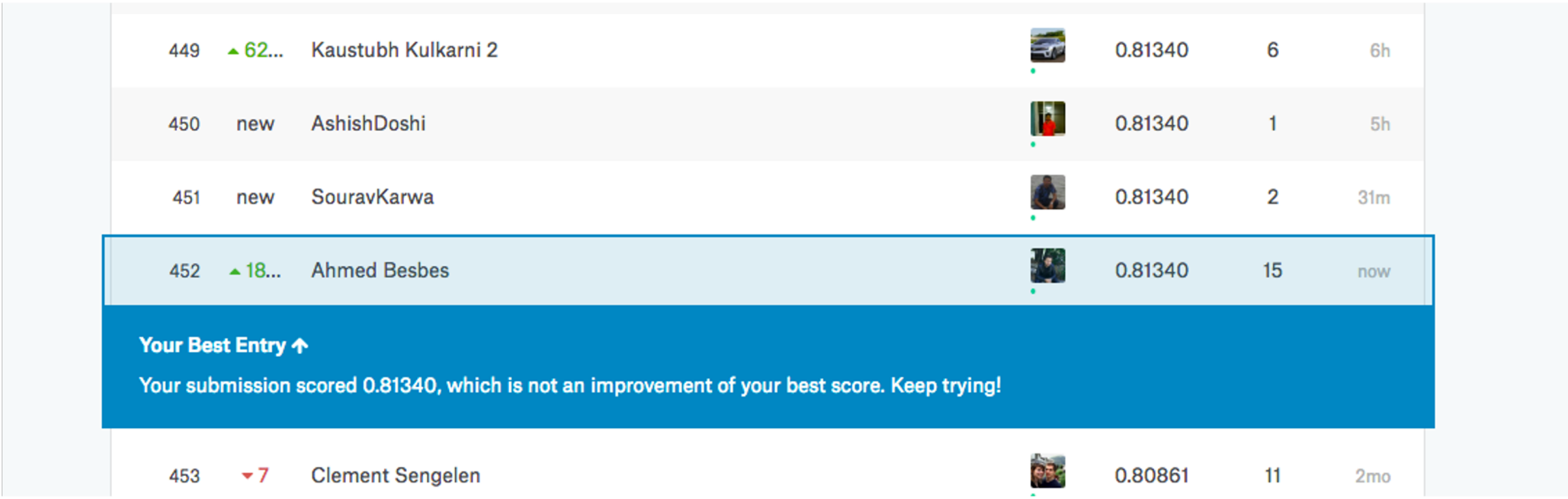
When to stop the update?

- Consider the updates of Logistic regression as trying to reduce the bias of the model
  - As we keep updating, the model overfits more to the training data
- We want to stop when the error on the validation set increases*
  - More on this later
- Validation test: a separate set that is used to measure overfitting
  - Training set
  - Validation set
  - Test set



More tricks?

- http://ahmedbesbes.com/how-to-score-08134-in-titanic-kaggle-challenge.html
- Feature Engineering/selection
- Parameter tuning
- Try different models



Distribution parameter estimation

- P(head) = θ, θ = #heads/#tosses
- HHTTH

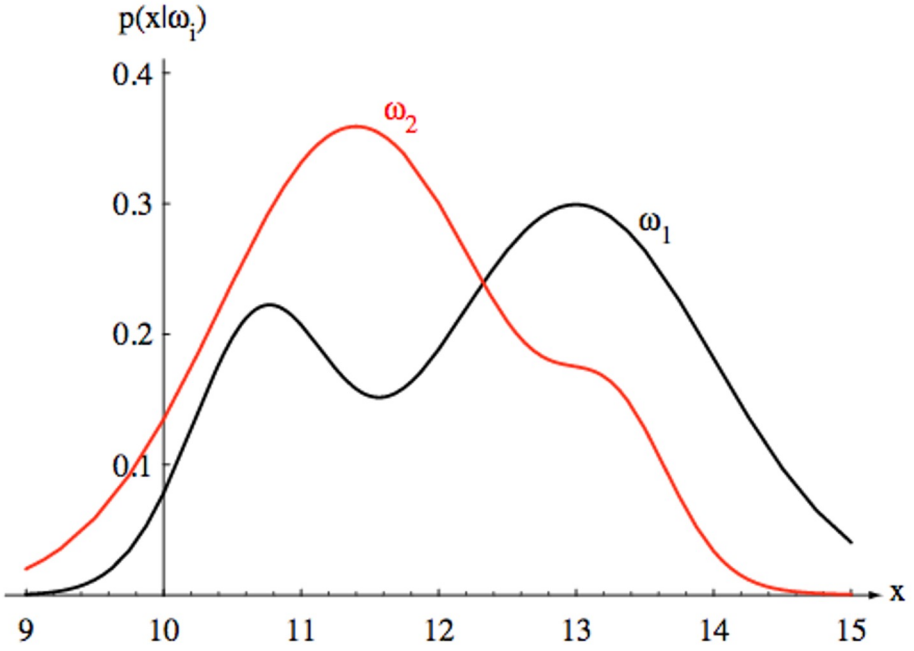
- L(θ) = P(X; θ) = P(HHTTH; θ)
- Maximum Likelihood Estimate (MLE)
  - Likelihood - Probability of encountering the data X given the parameters θ

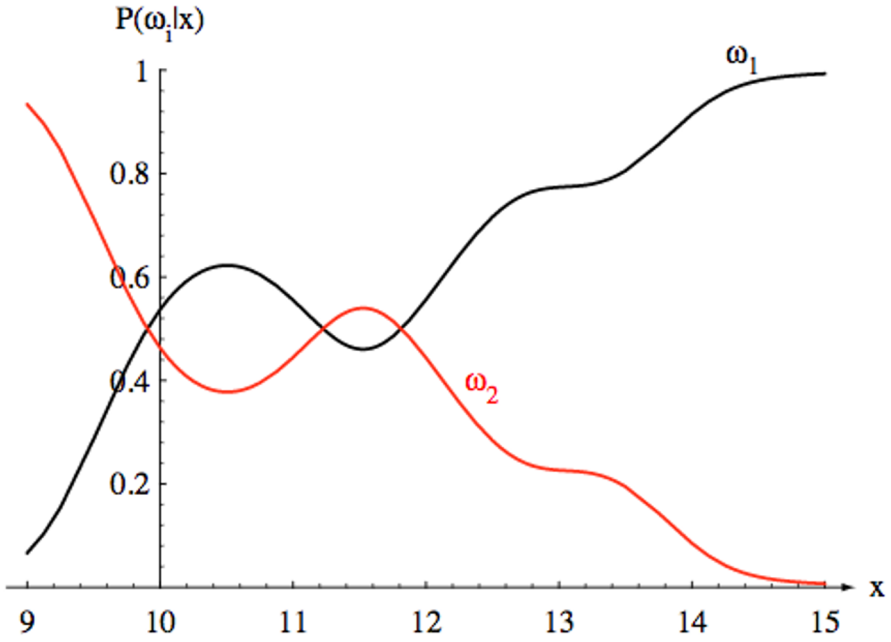
The Bayes Lecture

- Bayes Decision Rule
- Naïve Bayes

A simple decision rule

- If we can know either p(x|w) or p(w|x) we can make a classification guess

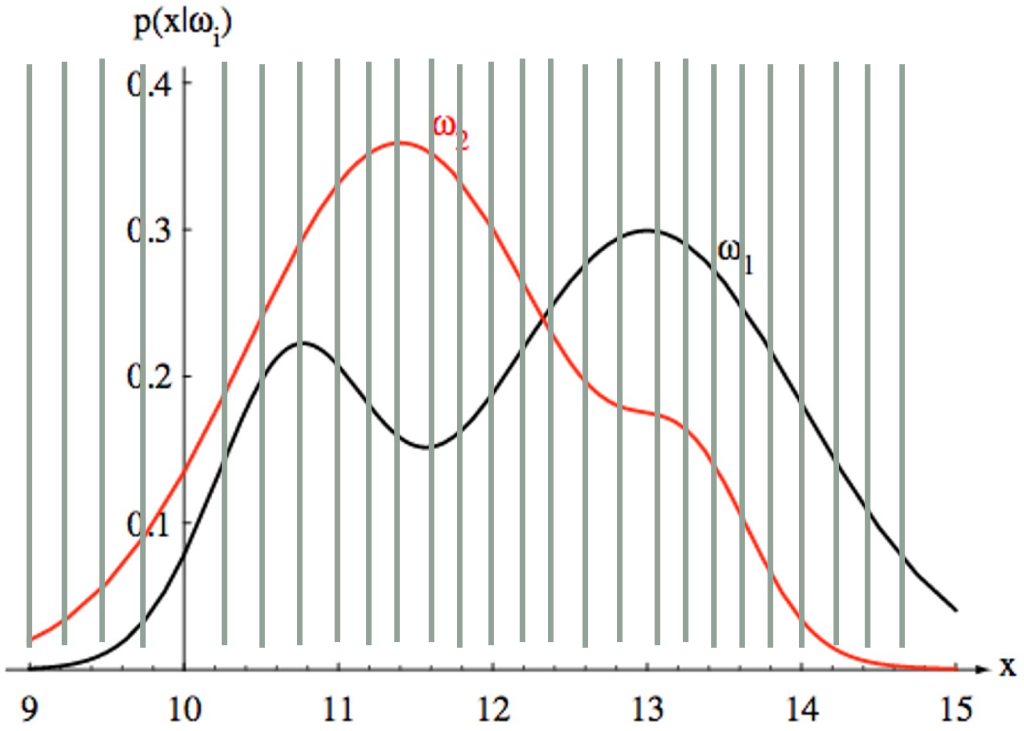




Goal: Find p(x|w) or p(w|x)

Duda et al. Pattern Classification

A simple way to estimate p(x|w)



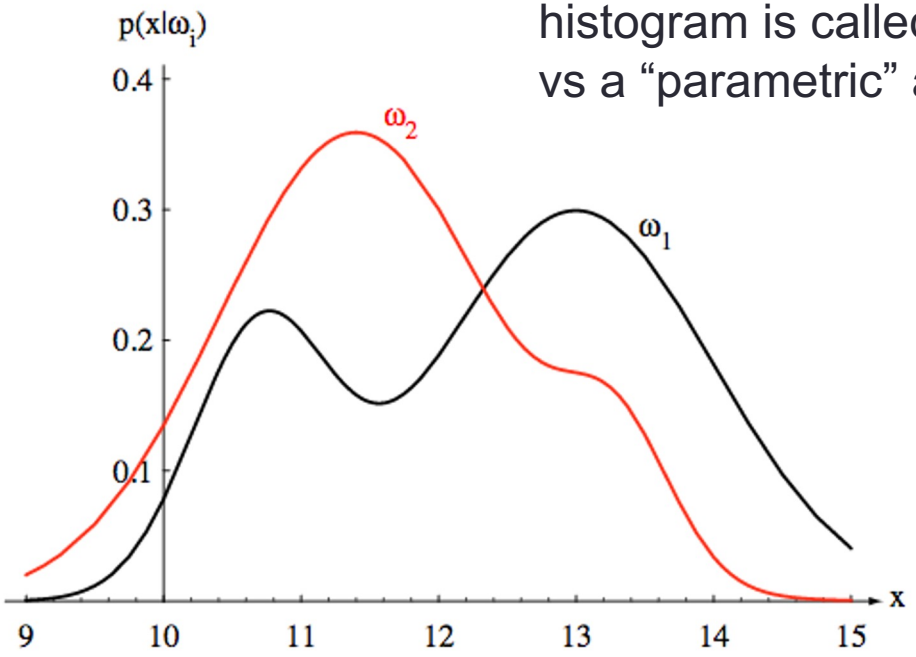
Make a histogram!

What happens if there is no data in a histogram bin?

The parametric approach

- We assume p(x|w) or p(w|x) follow some distributions with parameter θ

The method where we model the distribution using a histogram is called a “non-parametric” approach vs a “parametric” approach



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

Goal: Find θ so that we can estimate p(x|w) or p(w|x)

Maximum Likelihood Estimate (MLE)

$$p(w_i|x) = \frac{p(x|w_i)p(w_i)}{p(x)}$$

$$\text{Posterior} = \frac{\text{likelihood} * \text{prior}}{\text{evidence}}$$

- Maximizing the likelihood (probability of data given model parameters)

$p(\mathbf{x}|\theta) = L(\theta)$ <- This assumes the data is fixed

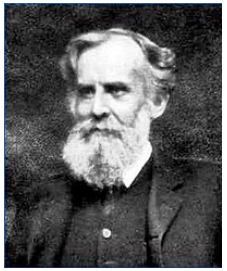
- Usually done on log likelihood

- Take the partial derivative wrt to θ and solve for the θ that maximizes the likelihood

MLE of Gaussian

- Observing $\{x_1, x_2, \ldots, x_i\}$, estimate the mean and the variance. Assume the data is normally distributed.

Frequentist vs Bayesian view



- Frequentist
  - Probability is “frequency of occurrence”
  - Data is from a random procedure that draw from unknown but fixed phenomenon.
    - Distribution parameter is a constant
- Bayesian
  - Probability is “degree of uncertainty”
  - Data is fixed and you want to infer about the unknown phenomenon.
    - Distribution parameter is a distribution
    - Prior knowledge about the phenomenon can change the inference results.

https://en.wikipedia.org/wiki/Frequentist_probability#/media/File:John_Venn.jpg
https://en.wikipedia.org/wiki/Thomas_Bayes#/media/File:Thomas_Bayes.gif

Maximum A Posteriori (MAP) Estimate

MLE

- Maximizing the likelihood (probability of data given model parameters)

$$\text{argmax}_{\theta}\ p(\mathbf{x}|\theta)$$

$$p(\mathbf{x}|\theta) = L(\theta)$$

- Usually done on log likelihood

- Take the partial derivative wrt to θ and solve for the θ that maximizes the likelihood

MAP

- Maximizing the posterior (model parameters given data)

$$\text{argmax}_{\theta}\ p(\theta|\mathbf{x})$$

- But we don't know $p(\theta|\mathbf{x})$

- Use Bayes rule

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

- Taking the argmax for θ we can ignore $p(\mathbf{x})$

- $\text{argmax}_{\theta}\ p(\mathbf{x}|\theta)\ p(\theta)$

MAP on Gaussian

- We know x is Gaussian with an unknown mean μ that we need to estimate and a known variance σ^2
- Assume the prior of μ is $N(\mu_0, \sigma_0^2)$

- MAP estimate of μ is

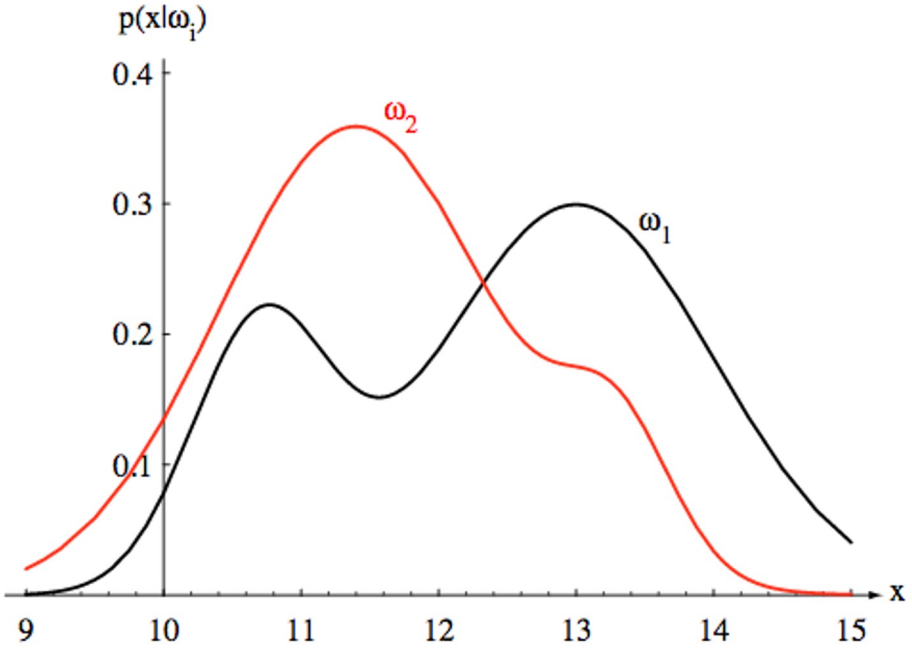
$$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right)\left[\frac{1}{n}\Sigma_{i=1}^{n} x_i\right] + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

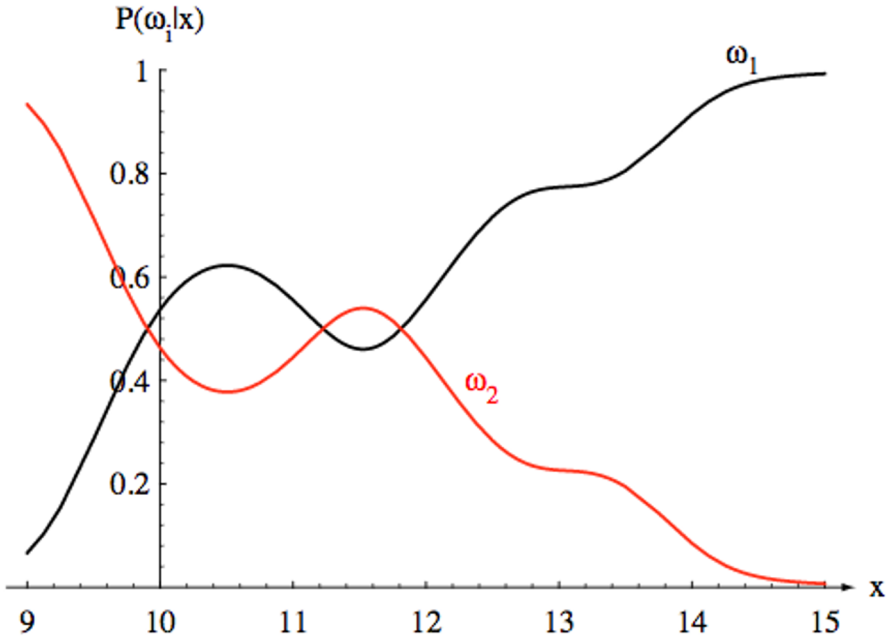
Notes of MAP estimate

- Usually harder to estimate than MLE
- If we use an uninformative prior for θ
  - MAP estimate = MLE
- Given infinite data
  - MAP estimate converges to MLE
- MAP is useful when you have less data, so you need additional knowledge about the domain
  - MAP estimate tends to converges to faster than MLE even with an arbitrary distribution
  - Can help prevent overfitting
- Useful for model adaptation (MAP adaptation)
  - Learn MLE on larger dataset, use this as your prior distribution
  - Learn MAP estimate on your dataset

A simple decision rule

- If we can know either p(x|w) or p(w|x) we can make a classification guess





Goal: Find p(x|w) or p(w|x) by finding the parameter of the distribution

Likelihood ratio test

- If $P(w_1|x) > P(w_2|x)$, that x is more likely to be class w_1
- Again we know $P(x|w_1)$ is more intuitive and easier to calculate than $P(w_1|x)$

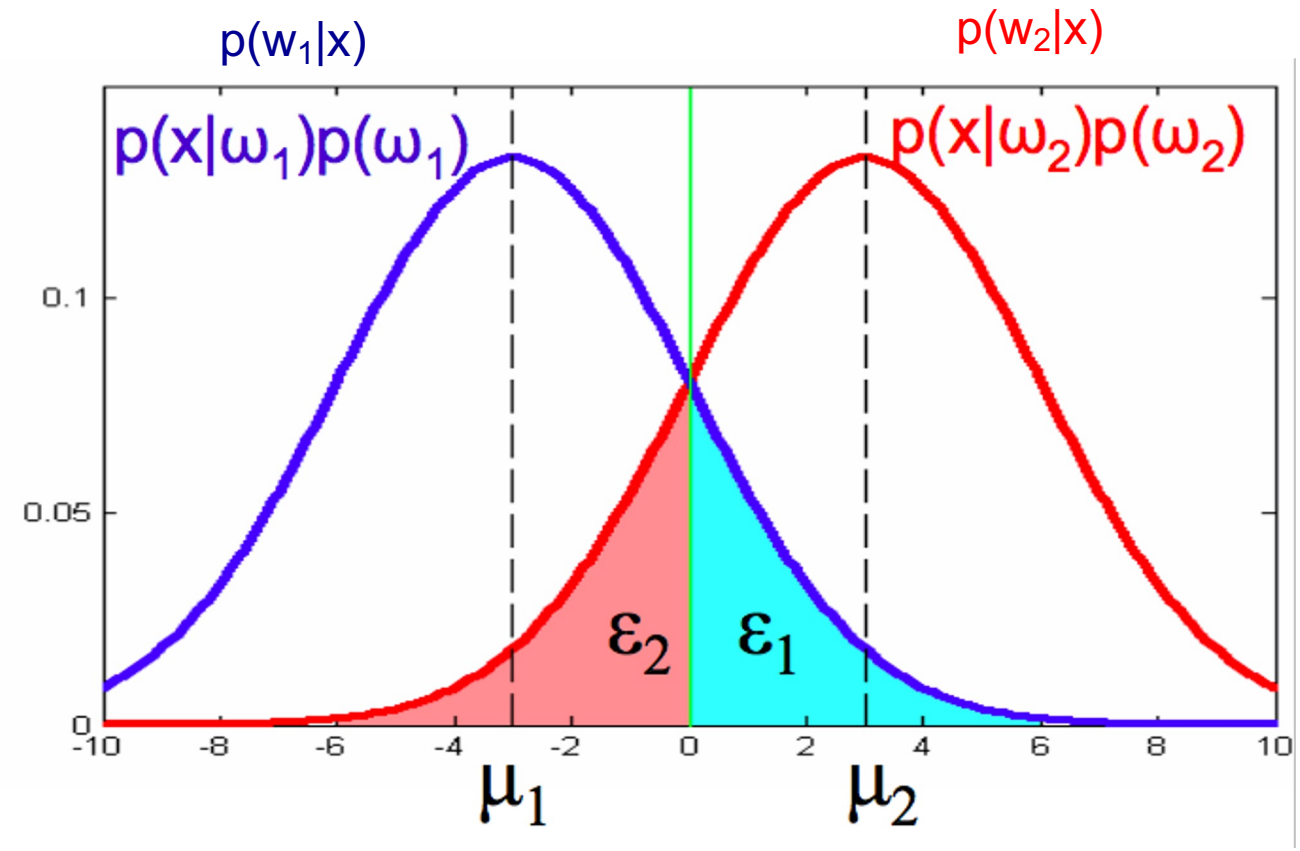
- Our classifier becomes
- $$P(x|w_1)P(w_1) \quad ? \quad P(x|w_2)P(w_2)$$
- $$\frac{P(x|w_1)}{P(x|w_2)} \quad ? \quad \frac{P(w_2)}{P(w_1)}$$

Likelihood ratio

Ratio of priors

Notes on likelihood ratio test (LRT)

- LRT minimizes the classification error (all errors are equally bad)

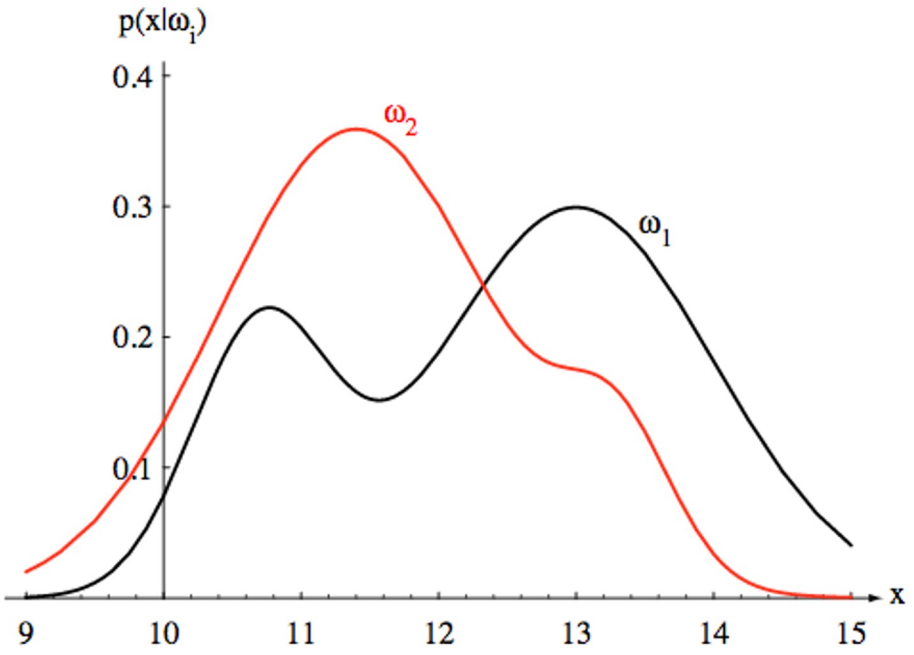


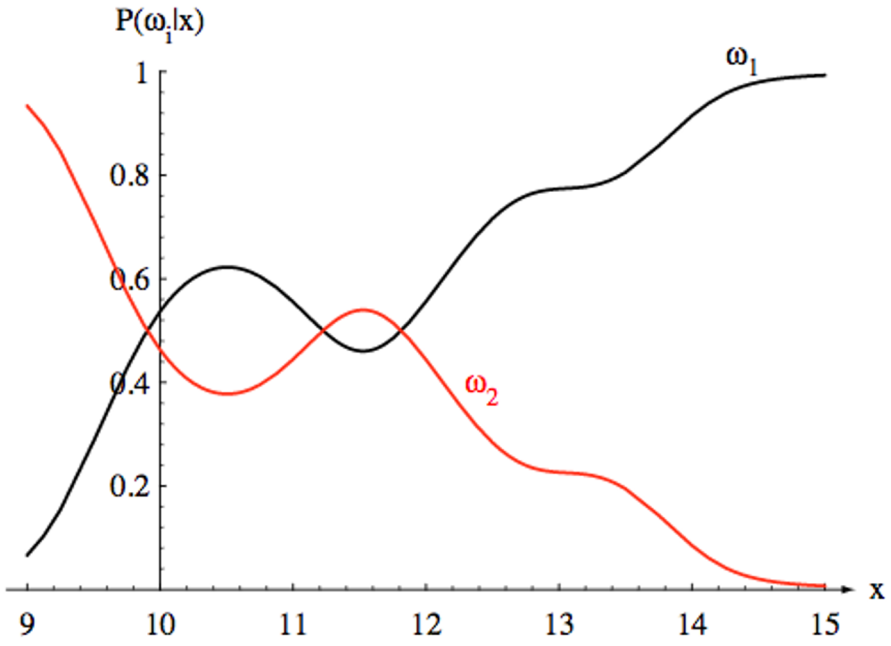
Notes on LRT

- If $P(w_1|x) > P(w_2|x)$, x is more likely to be class w_1
  - Also known as MAP decision rule
  - The classifier is sometimes called the Bayes classifier
- If we do not want to treat all error equally, we can assign different loss to each error, and minimize the expected loss. This is called Bayes loss/risk classifier
- $$\frac{P(x|w_1)}{P(x|w_2)} \quad ? \quad \frac{P(w_2)(L_{1|2} - L_{2|2})}{P(w_1)(L_{2|1} - L_{1|1})}$$
- When we treat errors equally, we refer to the zero-one loss
- $L_{1|2} = 1, L_{2|2} = 0, L_{2|1} = 1, L_{1|1} = 0$

Notes on LRT

- If we treat the priors as equal, we get the maximum likelihood criterion
- $$\frac{P(x|w_1)}{P(x|w_2)} \quad ? \quad 1$$





Naïve Bayes

- Below is the LRT or the Bayes classifier

$$P(x|w_1)P(w_1) \quad ? \quad P(x|w_2)P(w_2)$$

- What about Naïve Bayes?

- Here x is a vector with m features $[x_1,x_2,\ldots x_m]$
- $P(x|w_i)$ is m+1 dimensional
  - Sometimes to hard to model, not enough data, overfit, *curse of dimensionality*, etc.
- Assumes $x_1,x_2,\ldots x_m$ independent given w_i (conditional independence)
  - What does this mean?

Modeling distributions

Wind in the morning $X \in \{Calm, Windy\}$

PM2.5 level in the afternoon $Y \in \{Low, Med, High\}$

$\text{argmax } P(Y \mid X) = \text{argmax } P(X \mid Y)\, P(Y)$

Day	X	Y
1	W	M
2	C	M
3	W	M
4	W	H
5	C	L
6	W	L
7	C	H
8	W	L

Modeling distributions

Wind in the morning $X \in \{Calm, Windy\}$

PM2.5 level in the afternoon $Y \in \{Low, Med, High\}$

argmax P(Y | X)

Day	X	Y
1	W	M
2	C	M
3	W	M
4	W	H
5	C	L
6	W	L
7	C	H
8	W	L

P(X, Y)	L	M	H
C	1/8	1/8	1/8
W	2/8	2/8	1/8

Joint distribution

P(Y | X)	L	M	H
C	1/3	1/3	1/3
W	2/5	2/5	1/5

Conditional distribution

Modeling distributions

Wind in the morning $X \in \{Calm, Windy\}$

PM2.5 level in the afternoon $Y \in \{Low, Med, High\}$

argmax P(Y | X)

Day	X	Y
1	W	M
2	C	M
3	W	M
4	W	H
5	C	L
6	W	L
7	C	H
8	W	L

Joint distribution

P(X, Y)	L	M	H
C			
W			

Total data
8

count(X,Y)	L	M	H
C	1	1	1
W	2	2	1

$$P(X, Y) = \frac{\text{Count}(X, Y)}{\text{Total count}}$$ is the Maximum Likelihood Estimate (MLE) of P(X,Y)

Modeling distributions

Wind in the morning $X \in \{Calm, Windy\}$

PM2.5 level in the afternoon $Y \in \{Low, Med, High\}$

argmax P(Y | X)

Day	X	Y
1	W	M
2	C	M
3	W	M
4	W	H
5	C	L
6	W	L
7	C	H
8	W	L

P(Y | X)	L	M	H
C			
W			

count(X,Y)	L	M	H	Total
C	1	1	1	3
W	2	2	1	5

Conditional distribution

Total data
8

$$P(Y \mid X) = \frac{\text{Count}(X, Y)}{\text{Total count } (X)}$$ is the Maximum Likelihood Estimate (MLE) of P(Y|X)

Curse of dimensionality

Wind in the morning $X \in \{Calm, Windy\}$

PM2.5 level in the afternoon $Y \in \{Low, Med, High\}$

PM2.5 level in the evening $Z \in \{Low, Med, High\}$

$$\text{argmax } P(Z \mid Y, X) = \text{argmax } P(Y,X \mid Z)\, P(Z)$$

Day	X	Y	Z
1	W	L	M
2	C	M	M
3	W	H	M
4	W	M	H
5	C	M	L
6	W	M	L
7	C	L	H
8	W	H	L

count(Z,Y,X)	Z=L	Z=M	Z=H
X=W,Y=L	0	1	0
X=W,Y=M	1	0	1
X=W,Y=H	1	1	0
X=C,Y=L	0	0	1
X=C,Y=M	1	1	0
X=C,Y=H	0	0	0

Naïve Bayes

- $P(\mathbf{x}|w_i)P(w_i) = P(w_i) \prod_j P(x_j|w_i)$

- This assumption simplifies the calculation

Simplifying assumptions

Wind in the morning $X \in \{Calm, Windy\}$

PM2.5 level in the afternoon $Y \in \{Low, Med, High\}$

PM2.5 level in the evening $Z \in \{Low, Med, High\}$

$$\text{argmax } P(Z \mid Y, X) = \text{argmax } P(Y,X \mid Z)\, P(Z)$$
$$= \text{argmax } P(Y|Z)P(X|Z)P(Z)$$

Day	X	Y	Z
1	W	L	M
2	C	M	M
3	W	H	M
4	W	M	H
5	C	M	L
6	W	M	L
7	C	L	H
8	W	H	L

P(Y | Z)	Y = L	M	H
Z = L			
M			
H			

P(X | Z)	X = W	C
Z = L		
M		
H		

Dealing with zero probs

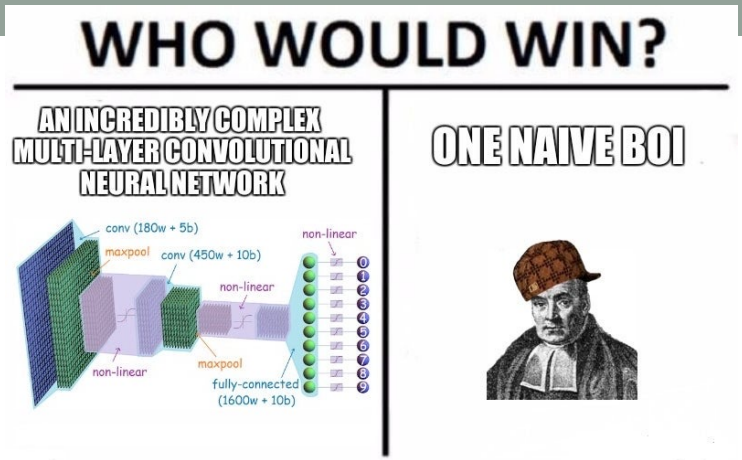
1. Use a very small value instead of zero (flooring)
2. Smooth the values using counts from other observations (smoothing)
3. Use priors (MAP adaptation)

Day	X	Y	Z
1	W	L	M
2	C	M	M
3	W	H	M
4	W	M	H
5	C	M	L
6	W	M	L
7	C	L	H
8	W	H	L

P(Y | Z)	Y = L	M	H
Z = L	0		
M			
H			

P(X | Z)	X = W	C
Z = L		
M		
H		

Naïve Bayes Notes

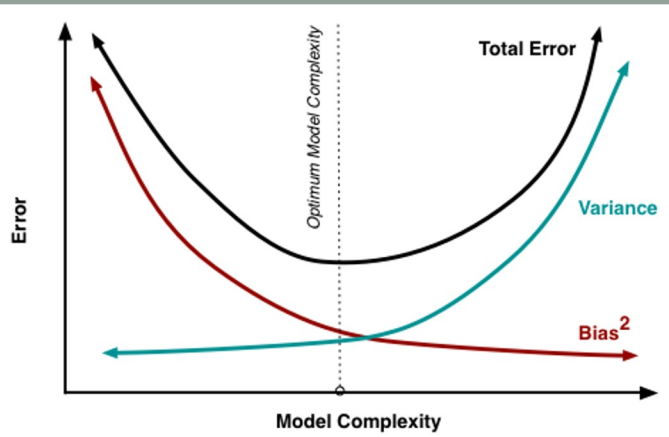


- $P(\mathbf{x}|w_i)P(w_i) = P(w_i) \prod_j P(x_j|w_i)$
- Note that we do not say anything about what kind of distribution $P(x_j|w_i)$ is.
  - In the homework you will play with this
    - Clean data
    - Estimate $P(x_j|w_i)$ using MLE, parametric and non-parametric version
    - Do prediction
    - Understand more about metrics
- Naïve Bayes can handle missing data
- Naïve Bayes is fast and quite good in practice
  - https://www.reddit.com/r/datascience/comments/hmhg9v/why_is_naive_bayes_so_popular_for_nlp/

Next homework

Summary

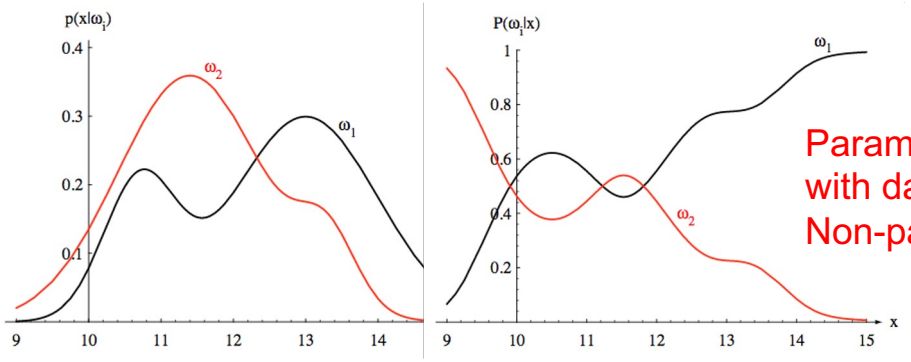
- Probabilistic view of linear regression
- Bias-Variance trade-off
  - Overfitting and underfitting
- MLE vs MAP estimate
  - How to use the prior
- LRT (Bayes Classifier)
  - Naïve Bayes



$$\frac{P(x|w_1)}{P(x|w_2)} \quad ? \quad \frac{P(w_2)}{P(w_1)}$$

Likelihood ratio

Ratio of priors

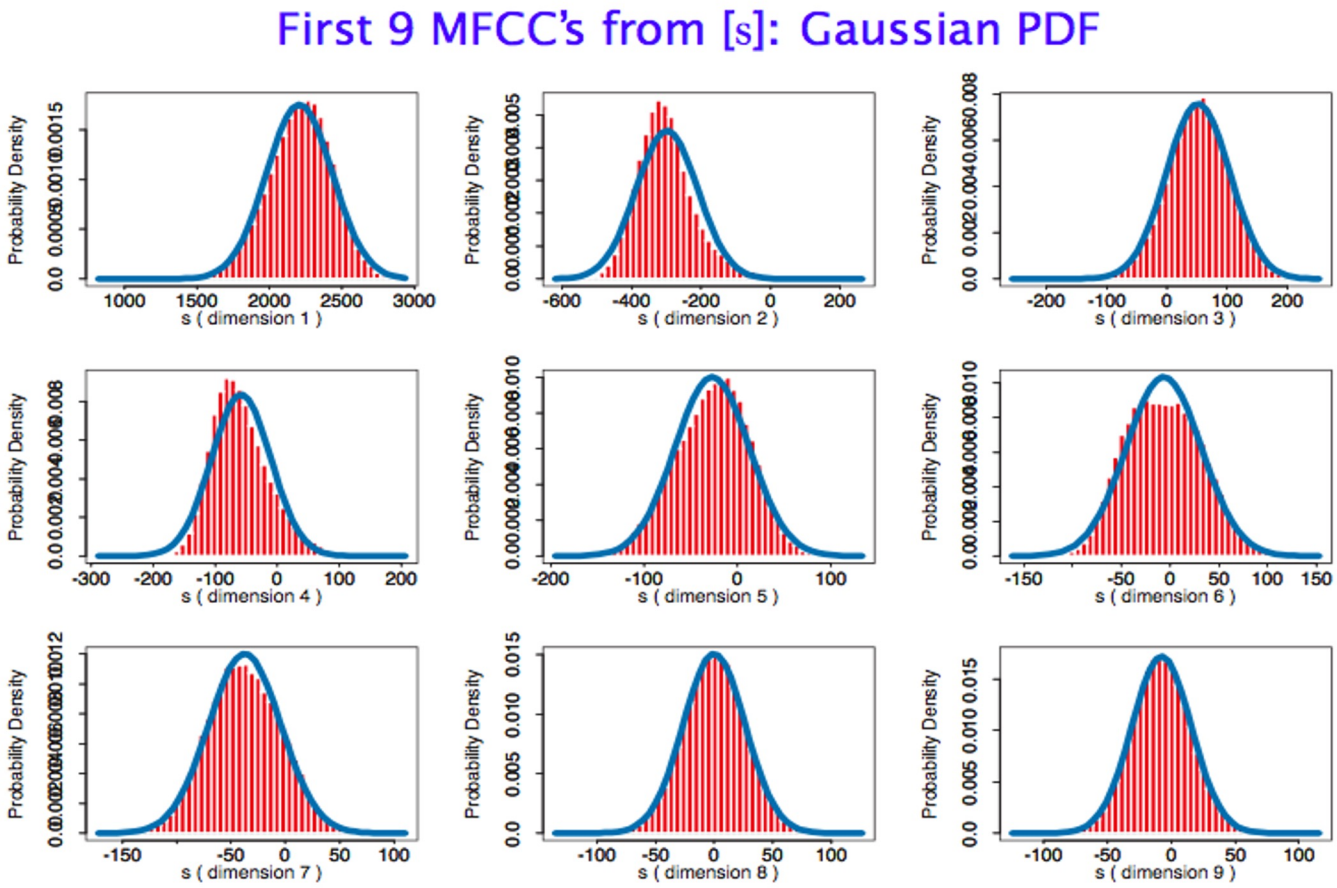


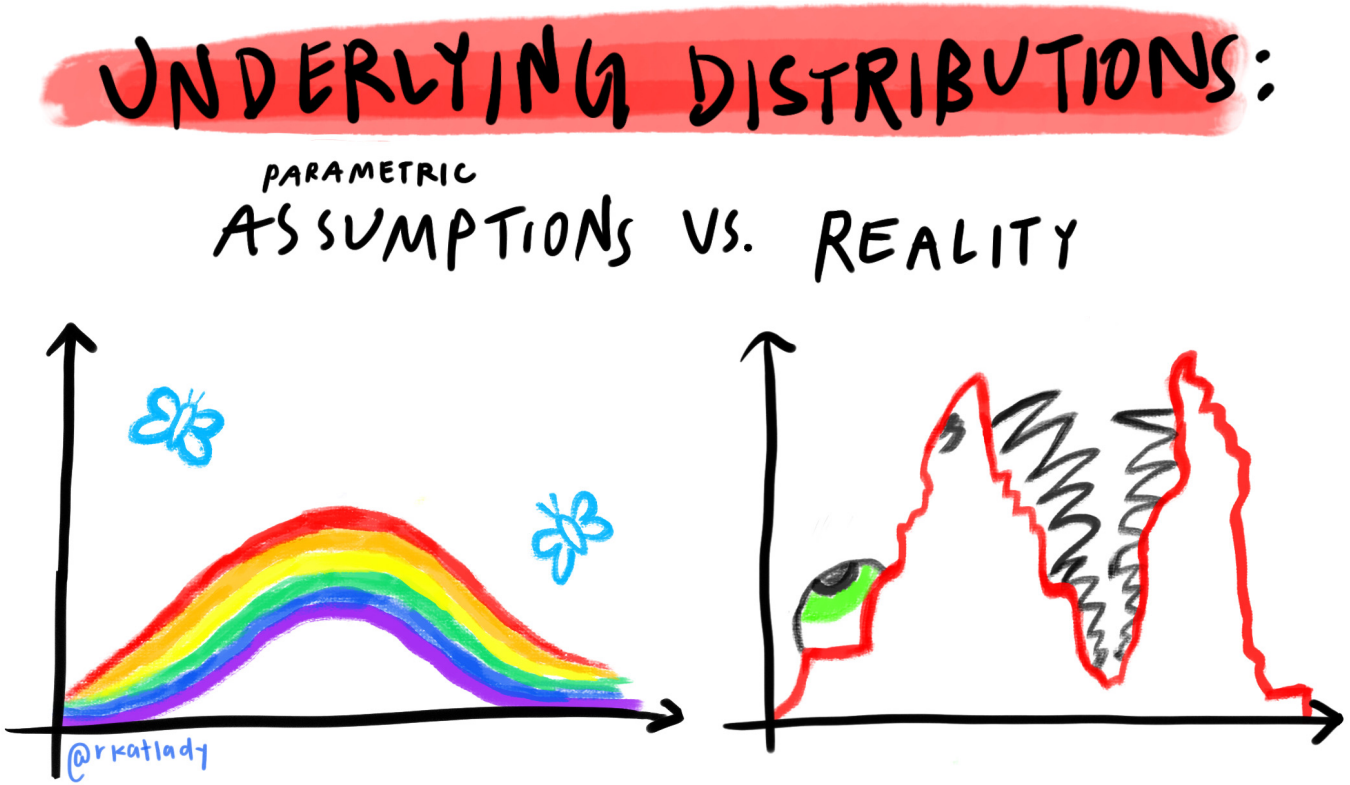
Parametric: need to assume the distribution (might not fit well with data)

Non-parametric: might encounter sparse bins

GMM

Can we do better than a single Gaussian?



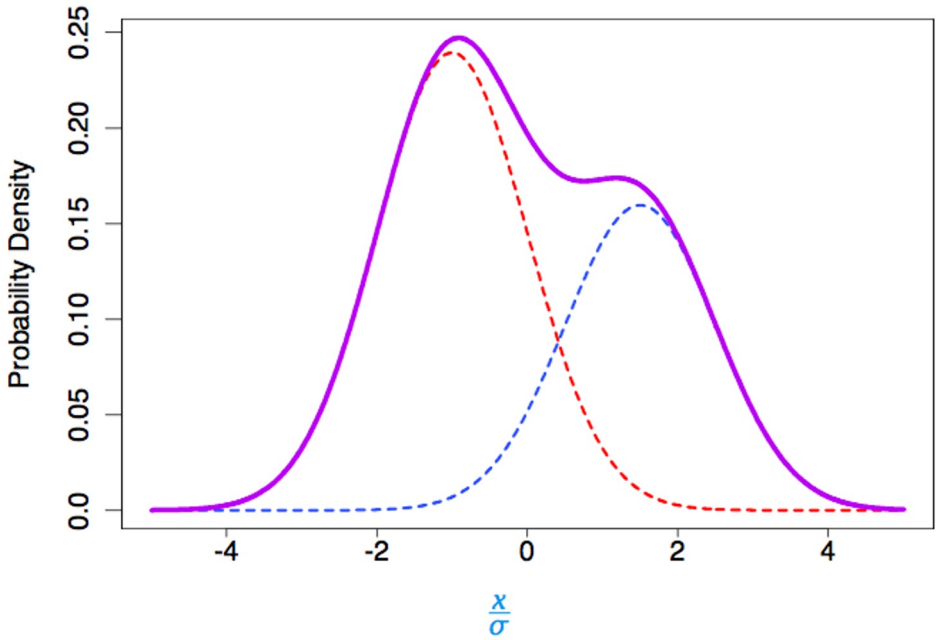


https://www.khstats.com/blog/tmle/tutorial/

Gaussian Mixture Models (GMMs)

- Gaussians cannot handle multi-modal data well
- Consider a class can be further divided into additional factors
- Mixing weight makes sure the overall probability sums to 1

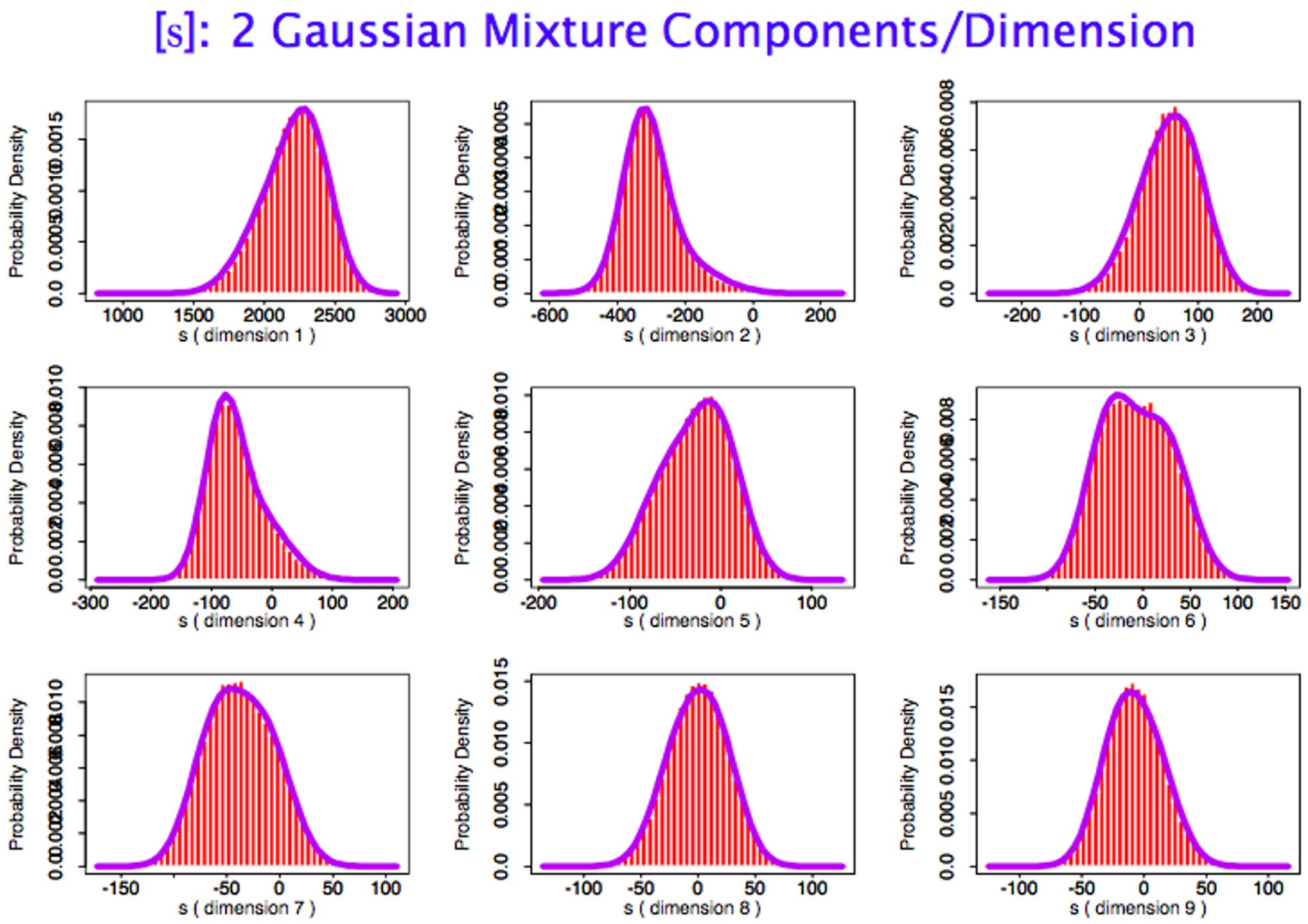
$$P(x) \sim \sum_{k=1}^{K} w_k N(\mu_k, \sigma_k)$$



$$p(x) = 0.6p_1(x) + 0.4p_2(x)$$

$$p_1(x) \sim N(-\sigma, \sigma^2) \qquad p_2(x) \sim N(1.5\sigma, \sigma^2)$$

Mixture of two Gaussians

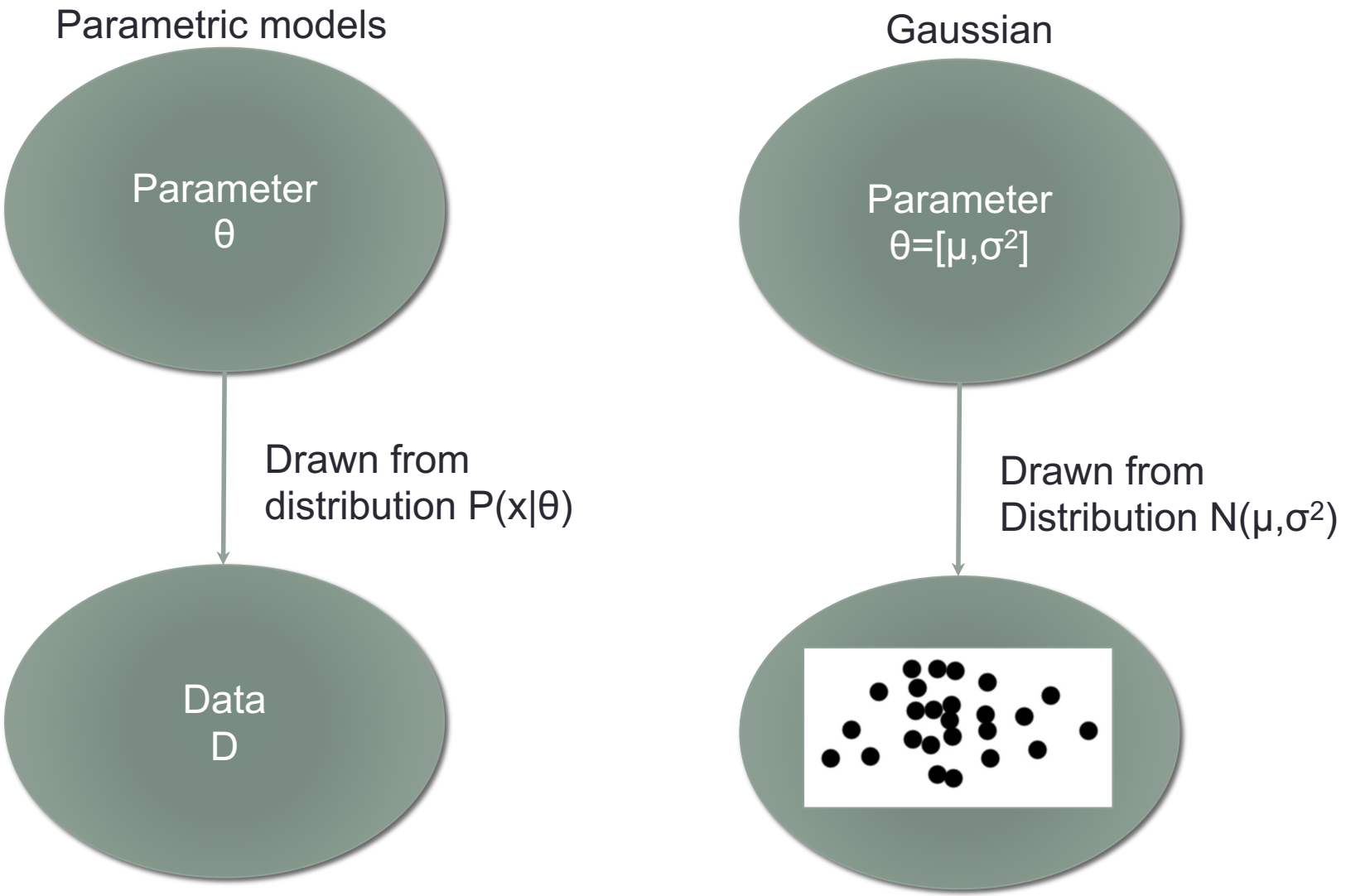


Mixture models

$$p(x) = \sum_k p(k) p_k(x)$$

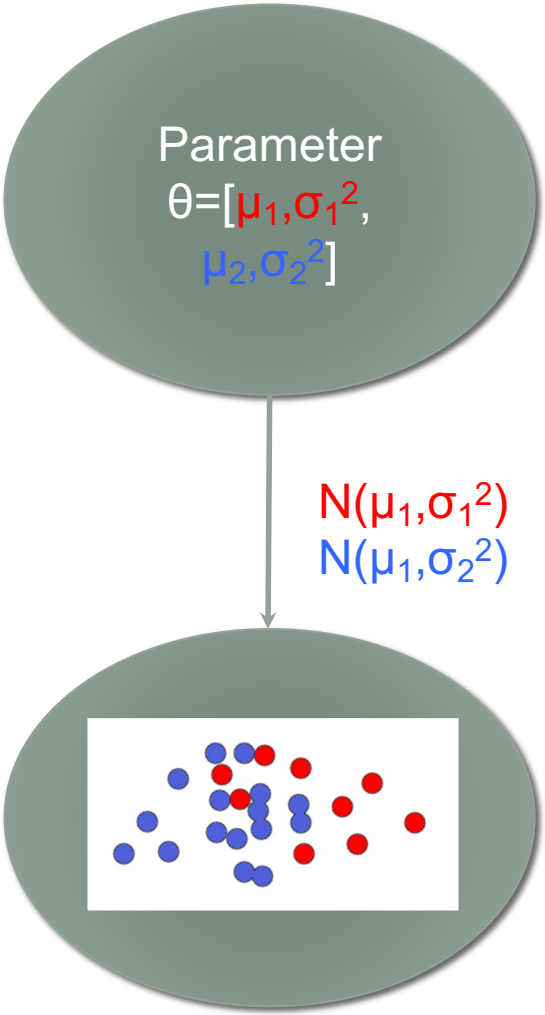
- A mixture of models from the same distributions (but with different parameters)
- Different mixtures can come from different sub-class
  - Cat class
    - Siamese cats
    - Persian cats
- p(k) is usually categorical (discrete classes)
- Usually the exact class for a sample point is unknown.
  - Latent variable

Parametric models

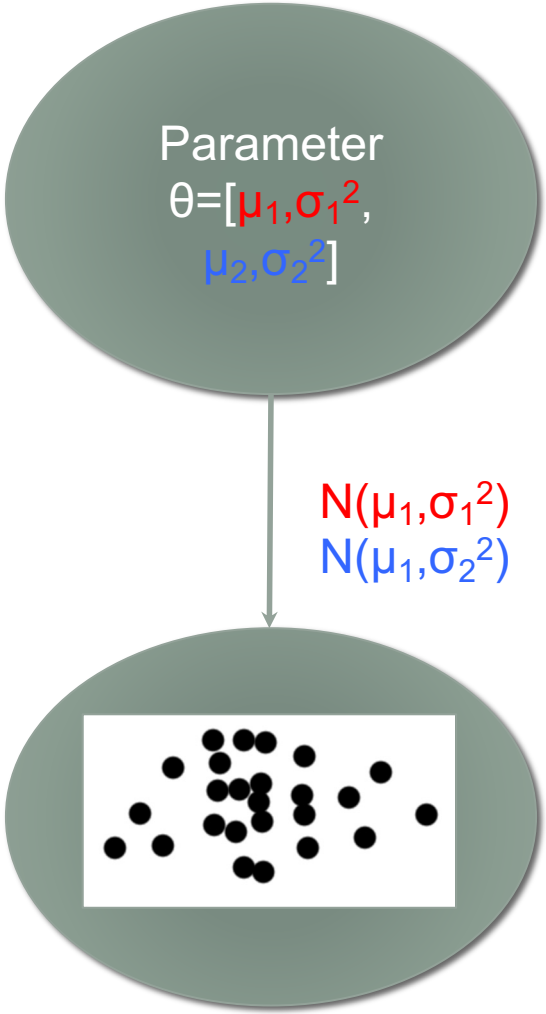


What if some data is missing?

Mixture of Gaussian

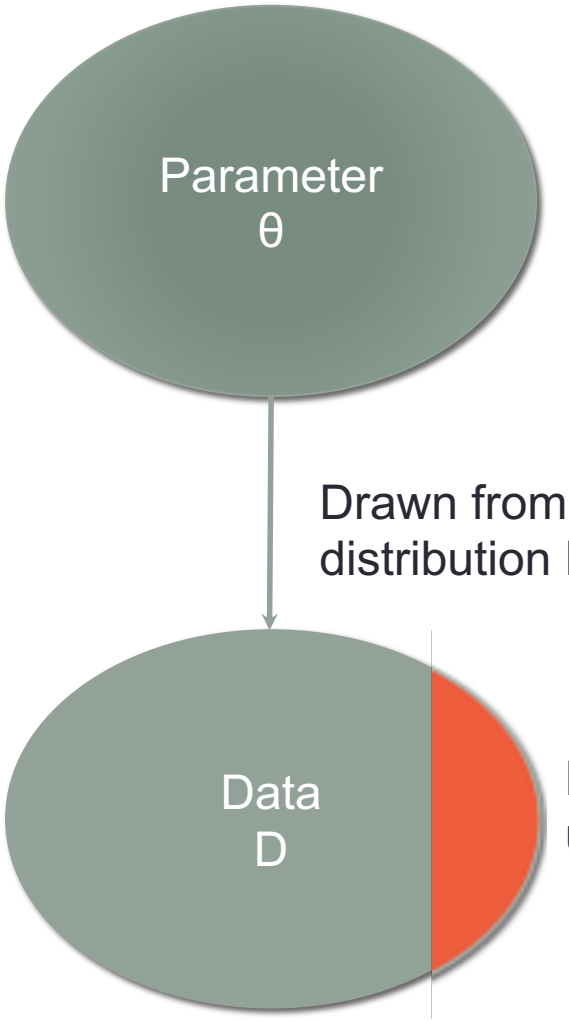


Unknown mixture labels



Estimating missing data

Parametric models



Need to estimate both the latent
Variables and the model parameters.

Slight difference in notation

$p(\mathbf{x}|\theta)$ vs $p(\mathbf{x};\theta)$

θ as a RV at a fixed value vs θ as a fixed parameter

Most of the time can be used interchangeably

Estimating latent variables and model parameters

- GMM $p(x) = \sum_k p(k) N(\mu_k, \sigma_k)$
- Observed ($x_1, x_2, \ldots, x_N$)
- Latent ($k_1, k_2, \ldots, k_N$) from K possible mixtures
- Parameter for p(k) is ϕ , $p(k = 1) = \phi_1$, $p(k = 2) = \phi_2 \ldots$

$$l(\phi, \mu, \Sigma) = \Sigma_{n=1}^{N} log p(x^{(i)}; \phi, \mu, \sigma)$$

$$= \Sigma_{n=1}^{N} log \Sigma_{l=1}^{K} p(x_n | k_{n,l}; \mu, \sigma) p(k_{n,l}; \phi)$$

Make things hard to solve

Cannot be solved by differentiating

Assuming k

- What if we somehow know k_n?
- Maximizing wrt to ϕ, μ, σ gives

$$\phi_j = \frac{1}{N}\Sigma_{n=1}^{N} 1(k_n = j)$$

$$\mu_j = \frac{\Sigma_{n=1}^{N} 1(k_n = j)x_n}{\Sigma_{n=1}^{N} 1(k_n = j)}$$

$$\sigma_j^2 = \frac{\Sigma_{n=1}^{N} 1(k_n = j)(x_n - \mu_j)^2}{\Sigma_{n=1}^{N} 1(k_n = j)}$$

$1(condition)$ Indicator function. Equals one if condition is met. Zero otherwise

Iterative algorithm

- Initialize ϕ, μ, σ
- Repeat till convergence
  - Expectation step (E-step) : Estimate the latent labels **k**
  - Maximization step (M-step) : Estimate the parameters ϕ, μ, σ given the latent labels
- Called Expectation Maximization (EM) Algorithm
- How to estimate the latent labels?

Iterative algorithm

- Initialize ϕ, μ, σ
- Repeat till convergence
  - Expectation step (E-step) : Estimate the latent labels **k** by finding the pdf of k given everything else $p(k| x ; \phi, \mu, \sigma)$
  - Maximization step (M-step) : Estimate the parameters ϕ, μ, σ given the latent labels by maximizing the expectation of the log likelihood
- Extension of MLE for latent variables
  - MLE : argmax log $p(x;\theta)$
  - EM : argmax log Σ_k $p(x, k;\theta)$

How to evaluate log Σ_k $p(x, k;\theta)$ when we don't know k?

Convex functions and Jensen's inequality

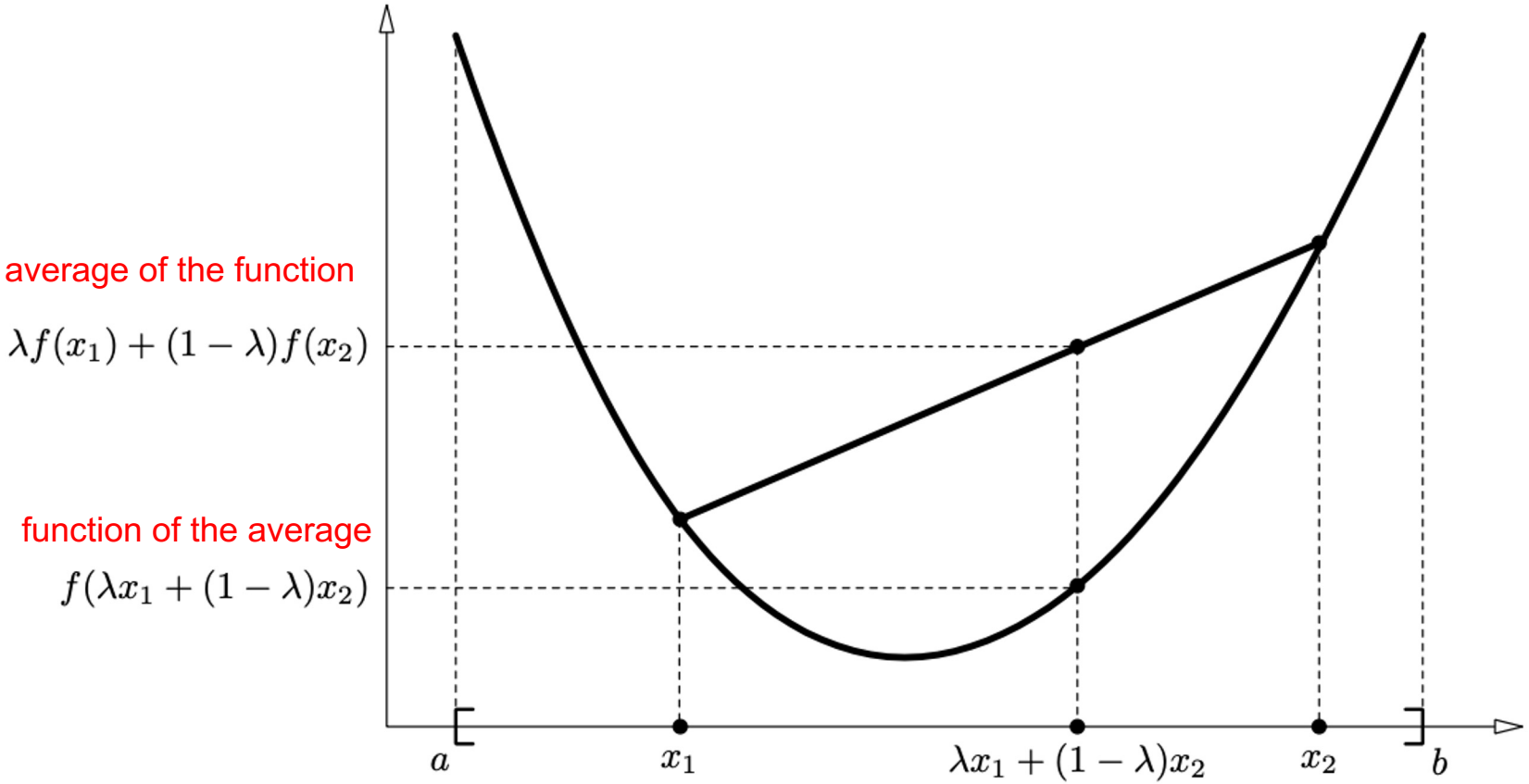


Figure 1: f is *convex* on $[a, b]$ if $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ $\forall x_1, x_2 \in [a, b], \quad \lambda \in [0, 1]$.

Jensen's inequality

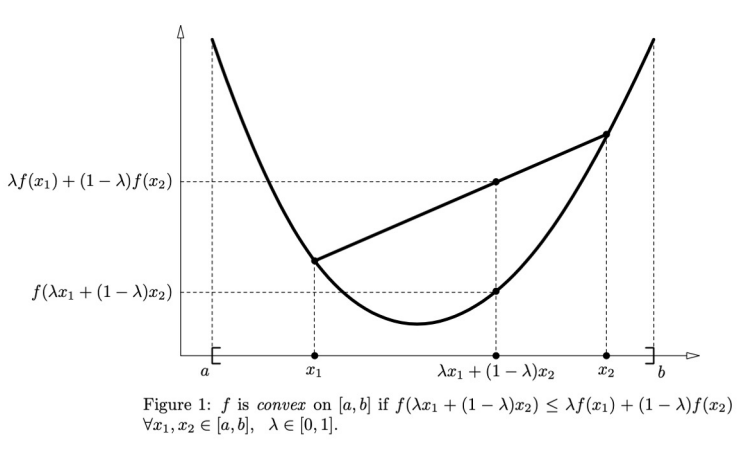


Figure 1: f is *convex* on $[a, b]$ if $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ $\forall x_1, x_2 \in [a, b], \quad \lambda \in [0, 1]$.

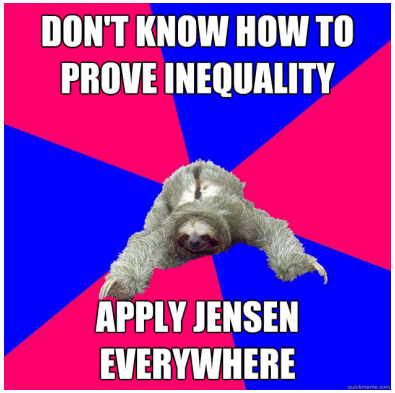
Let f be a convex function on interval I

If $x_1, x_2, \ldots, x_n$ is in I,

$w_1, \ldots, w_n > 0$ and sums to 1

then,

$$f\left(\sum_i^n w_i x_i\right) \leq \sum_i^n w_i f(x_i)$$



If f is concave, flip the inequality.

Can view this as expectation

$$f(E[X]) \leq E[f(X)]$$

Jensen's inequality and ELBO

log Σ_k p(x, k; θ)

$$f(\sum_i^n w_i x_i) \leq \sum_i^n w_i f(x_i)$$

Maximize Evidence Lower Bound (ELBO) = Σ_k Q(k) log (p(x, k;θ)/Q(k))

Making the lower bound tight

We will make the bound tight for fixed θ

Jensen's inequality is tight when?

$$f\left(\sum_i^n w_i x_i\right) \leq \sum_i^n w_i f(x_i)$$

$$f(E[X]) \leq E[f(X)]$$

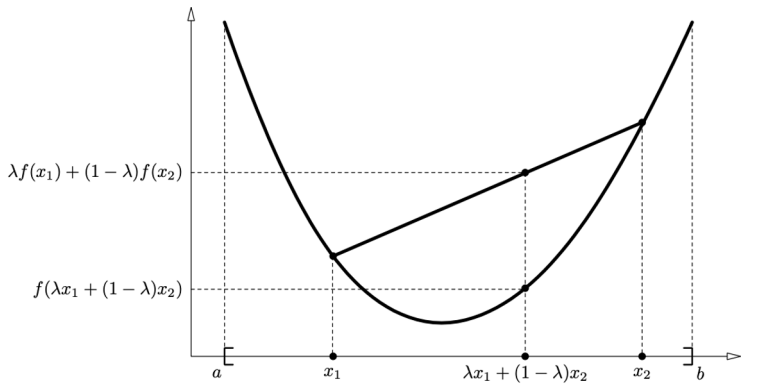


Figure 1: f is *convex* on $[a, b]$ if $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$ $\forall x_1, x_2 \in [a, b], \quad \lambda \in [0, 1]$.

Making the lower bound tight

We will make the bound tight for a fixed θ

$$f\left(\sum_i^n w_i x_i\right) \leq \sum_i^n w_i f(x_i)$$

$$f(E[X]) \leq E[f(X)]$$

If f( ) is strictly convex, Jensen's inequality is tight IFF

x_i are all equal

E[X] = X = constant

Making the lower bound tight

We will make the bound tight for a fixed θ

Jensen's inequality is tight when the inside of the expectation is a constant, c wrt the expectation of k

$p(x, k;\theta)/Q(k) = c$

or $Q(k) = p(k \mid x; \theta)$

Iterative algorithm (general)

- Goal of EM : $\log \sum_k p(x, k;\theta) >= \Sigma_k Q(k) \log (p(x, k;\theta)/Q(k))$
- Maximize the ELBO instead
- Initialize Θ
- Repeat till convergence
  - Expectation step (E-step) : estimate the conditional expectation $Q(k) = p(k|x;\theta)$ using the current θ.
  - Maximization step (M-step) : Estimate new Θ given by maximizing the ELBO given current Q(k)

EM on a simple example

- Grades in class P(A) = 0.5 P(B) = 0.5-θ P(C) = θ
- We want to estimate θ from three known numbers
  - N_a N_b N_c
- Find the maximum likelihood estimate of θ

EM on a simple example

- Grades in class P(A) = 0.5 P(B) = 0.5-θ P(C) = θ
- We want to estimate θ from ONE known number
  - N_c (we also know N the total number of students)
- Find θ using EM

Will this work?

For iteration i, with $\theta^{(i)}$

ELBO

$\log \sum_k p(x, k;\theta^{(i)}) >= \Sigma_k Q(k) \log (p(x, k;\theta^{(i)})/Q(k))$

E-step, making the bound tight by picking Q'(k) yields

$\log \sum_k p(x, k;\theta^{(i)}) = \Sigma_k Q'(k) \log (p(x, k;\theta^{(i)})/Q'(k))$

M-step, maximize ELBO by finding $\theta^{(i+1)}$

$\Sigma_k Q'(k) \log (p(x, k;\theta^{(i)})/Q'(k)) <= \Sigma_k Q'(k) \log (p(x, k;\theta^{(i+1)})/Q'(k))$

For iteration i+1, with $\theta^{(i+1)}$

$\log \sum_k p(x, k;\theta^{(i+1)}) >= \Sigma_k Q(k) \log (p(x, k;\theta^{(i+1)})/Q(k))$

Thus,

$\log \sum_k p(x, k;\theta^{(i+1)}) >= \log \sum_k p(x, k;\theta^{(i)})$

So EM improves the likelihood at every step!

Notes on ELBO

We set $Q(k) = p(k \mid x; \theta)$ to make the inequality tight.

What if we cannot compute $p(k \mid x; \theta)$?

Use a looser bound by picking any $Q(k)$

Estimate $p(k \mid x; \theta)$ with $q(k \mid x; \theta)$ that we can compute

This is called Variational Inference

We will revisit this.

Estimating latent variables and model parameters

- GMM $p(x) = \sum_k p(k) N(\mu_k, \sigma_k)$
- Observed ($x_1, x_2, \ldots, x_N$)
- Latent ($k_1, k_2, \ldots, k_N$) from K possible mixtures
- Parameter for p(k) is ϕ , $p(k = 1) = \phi_1$, $p(k = 2) = \phi_2 \ldots$

$$l(\phi, \mu, \Sigma) = \Sigma_{n=1}^{N} log p(x^{(i)}; \phi, \mu, \sigma)$$

$$= \Sigma_{n=1}^{N} log \Sigma_{l=1}^{K} p(x_n | k_{n,l}; \mu, \sigma) p(k_{n,l}; \phi)$$

Make things hard to solve

Cannot be solved by differentiating

EM on GMM

- E-step
  - Set soft labels: $w_{n,j}$ = probability that nth sample comes from jth mixture p
  - Using Bayes rule
    - $p(k|x ; \mu, \sigma, \phi) = \frac{p(x|k ; \mu, \sigma, \phi)\, p(k; \mu, \sigma, \phi)}{p(x; \mu, \sigma, \phi)}$
    - $p(k|x ; \mu, \sigma, \phi)$ is proportional to $p(x|k ; \mu, \sigma, \phi)\, p(k; \phi)$

$N(\mu_j, \sigma_j)$ ϕ_j

$$p(k_n = j|x_n; \phi, \mu, \Sigma) = \frac{p(x_n; \mu_j, \sigma_j)p(k_n = j; \phi)}{\Sigma_l p(x_n; \mu_l, \sigma_l)p(k_n = l; \phi)}$$

EM on GMM

- M-step (hard labels)

$$\phi_j = \frac{1}{N}\Sigma_{n=1}^{N} 1(k_n = j)$$

$$\mu_j = \frac{\Sigma_{n=1}^{N} 1(k_n = j)x_n}{\Sigma_{n=1}^{N} 1(k_n = j)}$$

$$\sigma_j^2 = \frac{\Sigma_{n=1}^{N} 1(k_n = j)(x_n - \mu_j)^2}{\Sigma_{n=1}^{N} 1(k_n = j)}$$

EM on GMM

- M-step (soft labels)

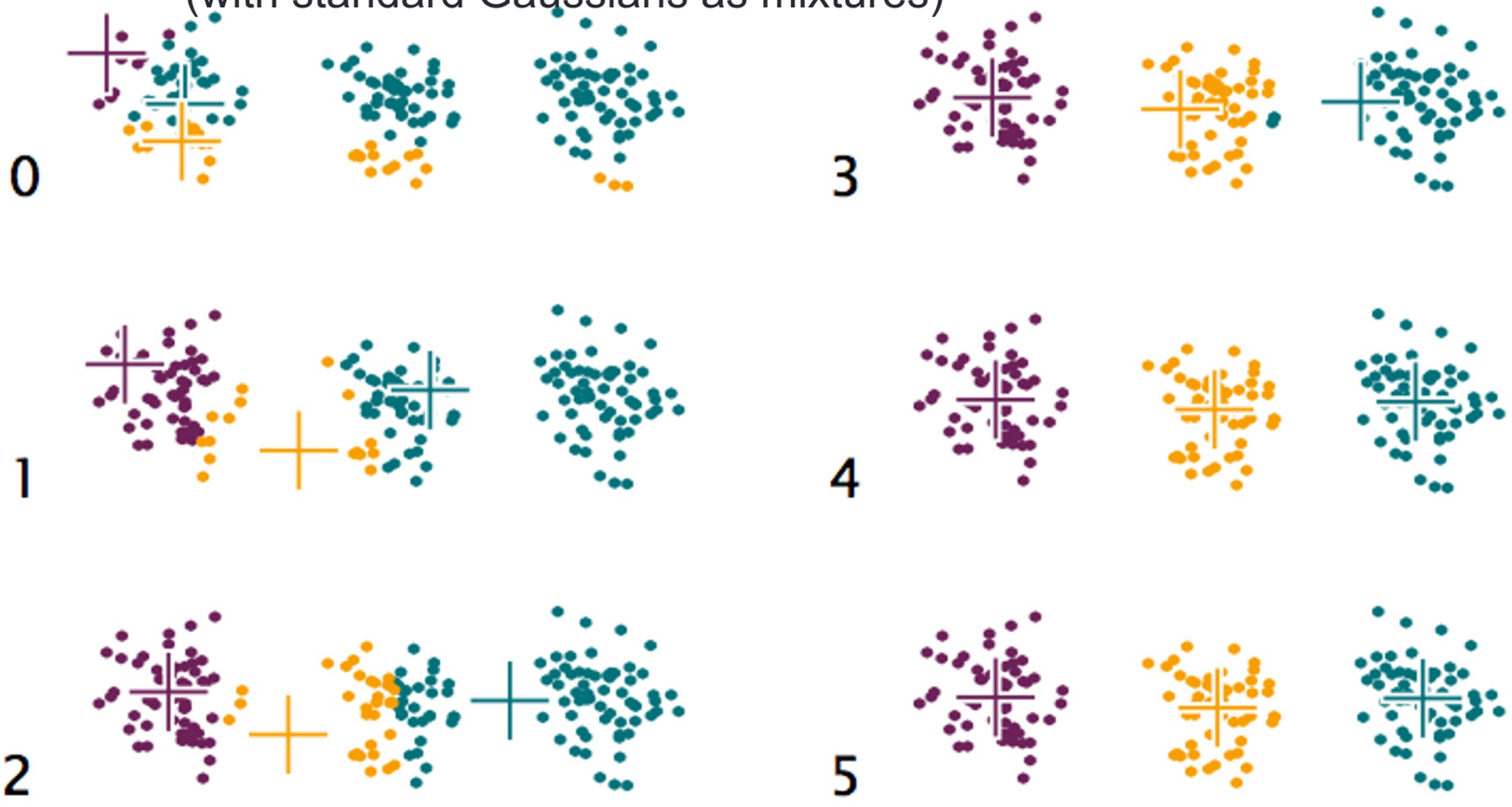
$$\phi_j = \frac{1}{N}\Sigma_{n=1}^{N} w_{n,j}$$

$$\mu_j = \frac{\Sigma_{n=1}^{N} w_{n,j} x_n}{\Sigma_{n=1}^{N} w_{n,j}}$$

$$\sigma_j^2 = \frac{\Sigma_{n=1}^{N} w_{n,j} (x_n - \mu_j)^2}{\Sigma_{n=1}^{N} w_{n,j}}$$

K-mean vs EM

EM on GMM can be considered as EM with soft labels
(with standard Gaussians as mixtures)



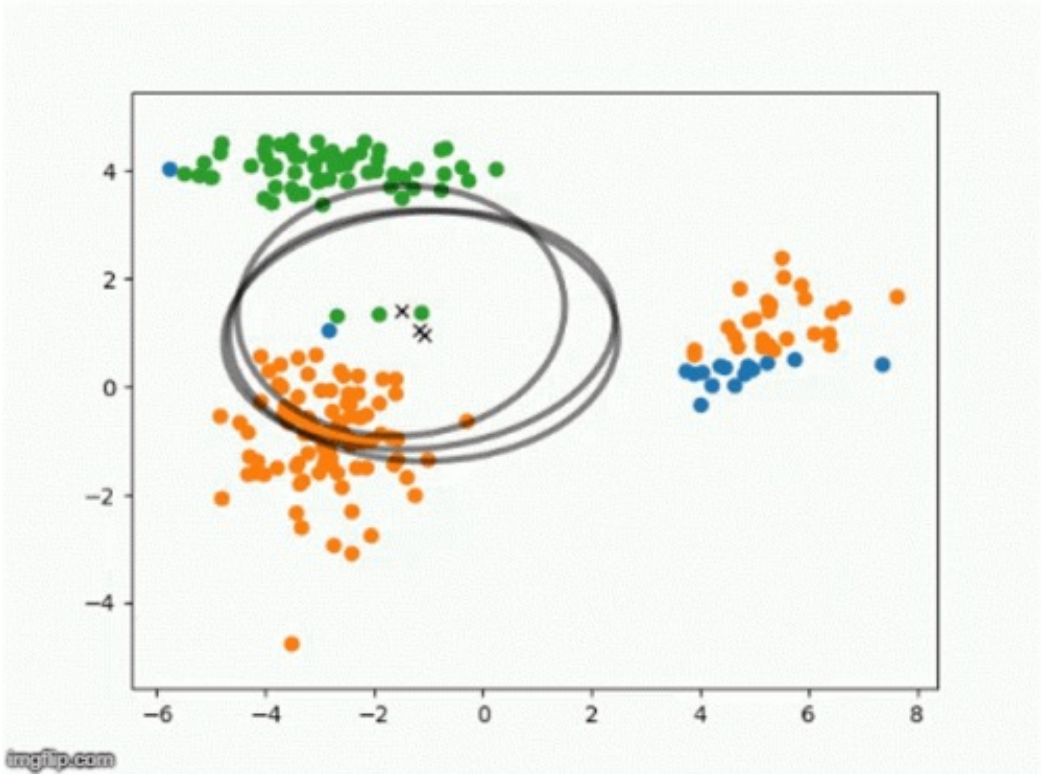
K-mean clustering

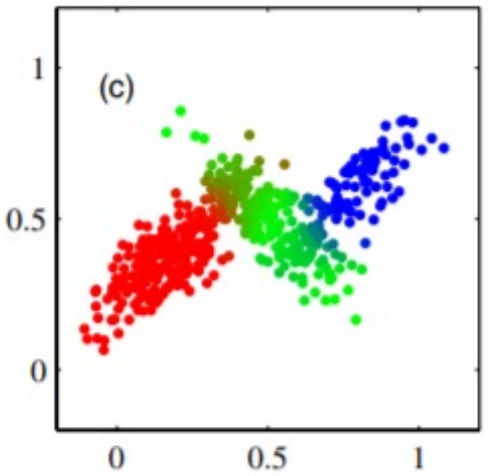
- Task: cluster data into groups
- K-mean algorithm
  - Initialization: Pick K data points as cluster centers
  - Assign: Assign data points to the closest centers
  - Update: Re-compute cluster center
  - Repeat: Assign and Update

EM algorithm for GMM

- Task: cluster data into Gaussians
- EM algorithm
  - Initialization: Randomly initialize parameters Gaussians
  - Expectation: Assign data points to the closest Gaussians
  - Maximization: Re-compute Gaussians parameters according to assigned data points
  - Repeat: Expectation and Maximization
- Note: assigning data points is actually a soft assignment (with probability)

K-mean vs EM

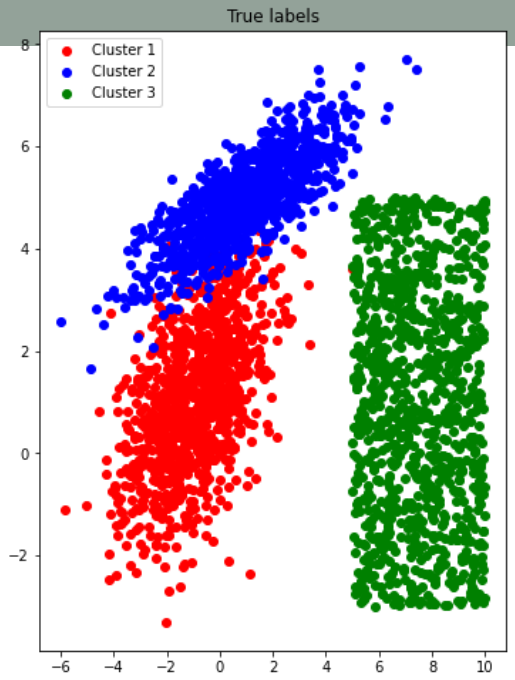


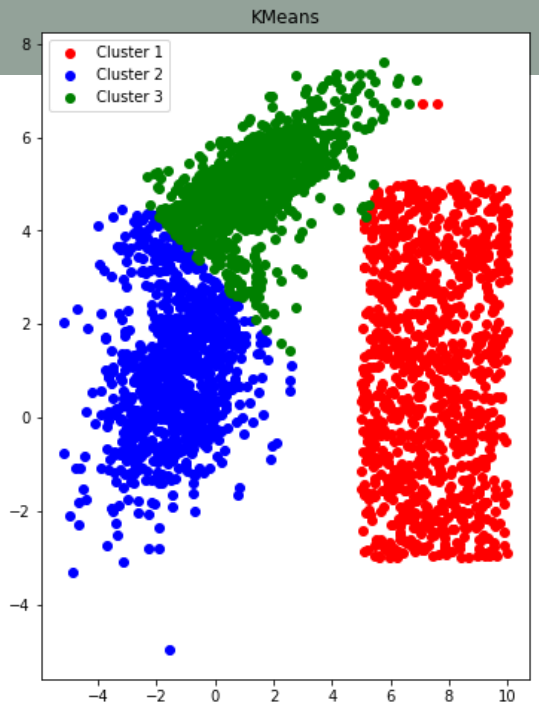


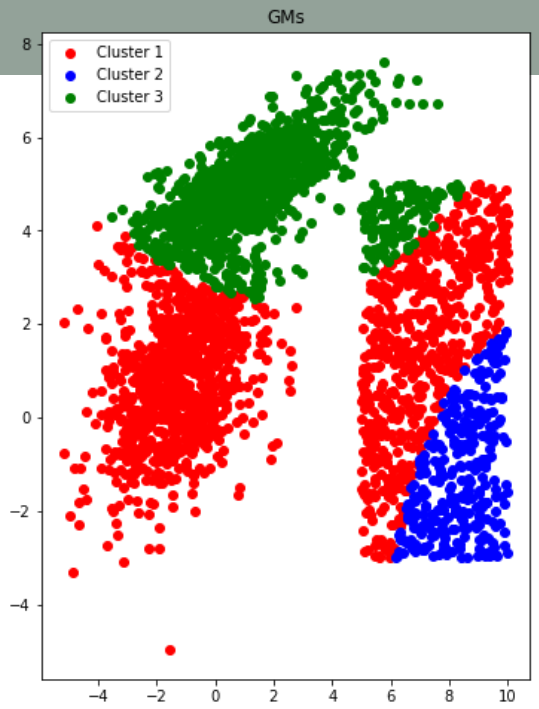
https://towardsdatascience.com/gaussian-mixture-models-vs-k-means-which-one-to-choose-62f2736025f0

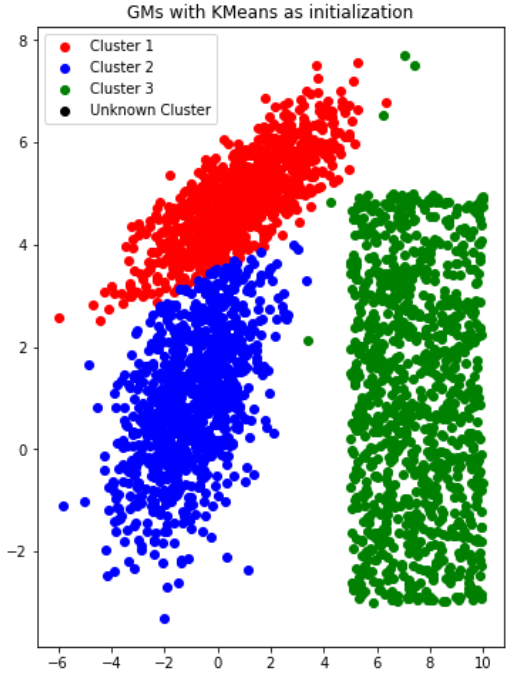
EM/GMM notes

- Converges to local maxima (maximizing likelihood)
  - Just like k-means, need to try different initialization points
- EM always improve the likelihood for each iteration
  - Stops EM when likelihood changes < threshold
- Just like k-means some centroid can get stuck with one sample point and no longer moves
  - For EM on GMM this cause variance to go to 0…
    - Introduce variance floor (minimum variance a Gaussian can have)
- Tricks to avoid bad local maxima
  - Starts with 1 Gaussian
  - Split the Gaussians according to the direction of maximum variance
  - Repeat until arrive at k Gaussians
  - Does not guarantee global maxima but works well in practice



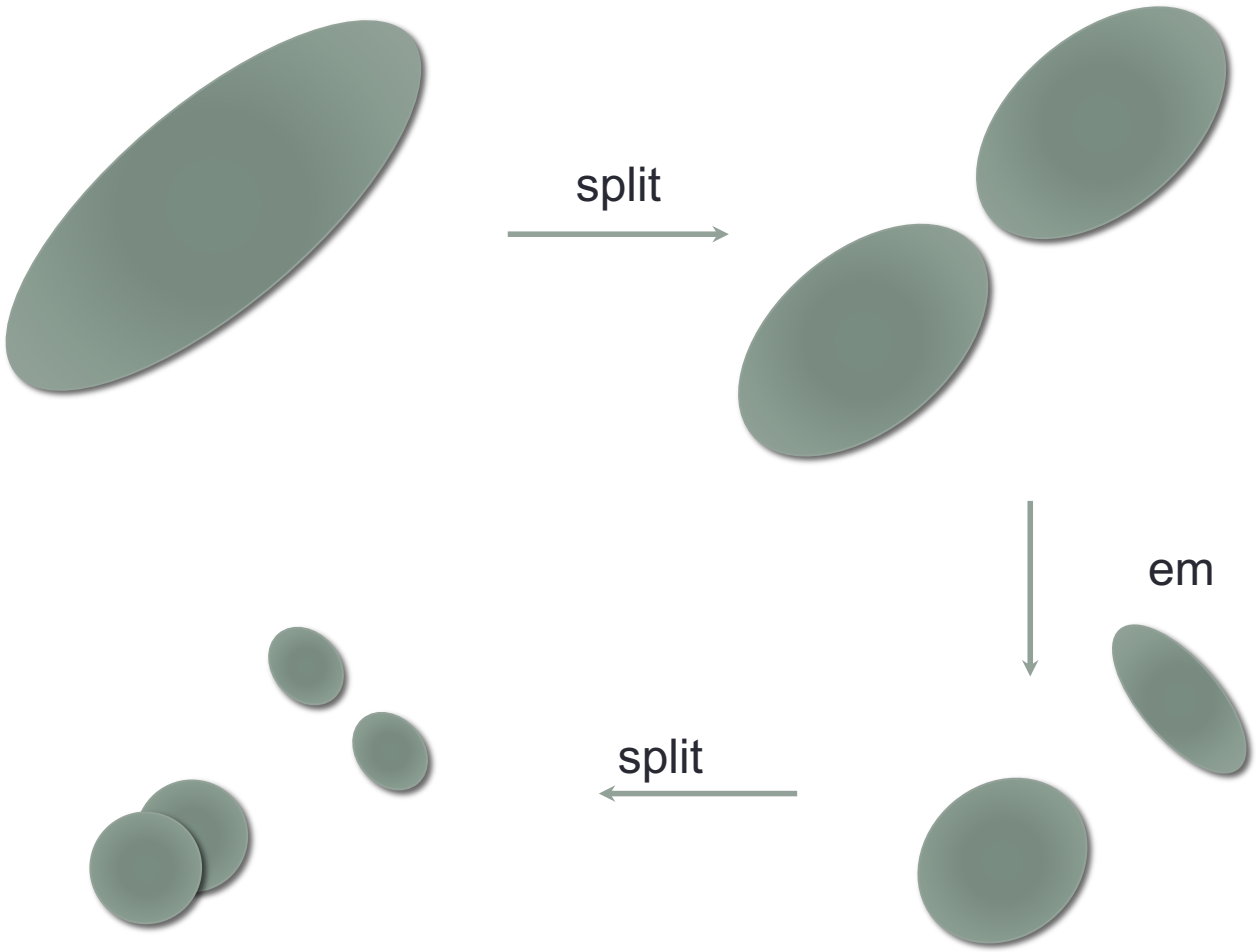






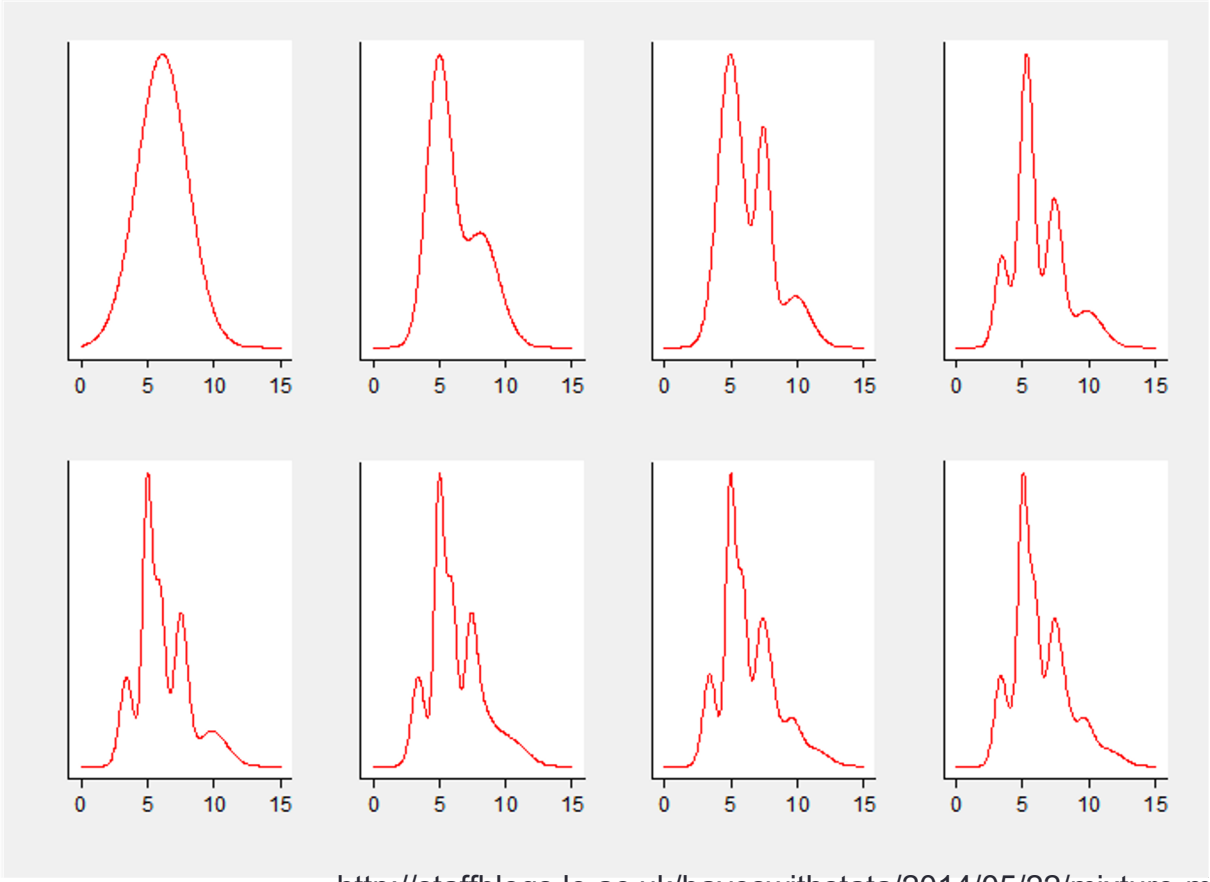
https://towardsdatascience.com/gaussian-mixture-models-vs-k-means-which-one-to-choose-62f2736025f0

Gaussian splitting



Picking the amount of Gaussians

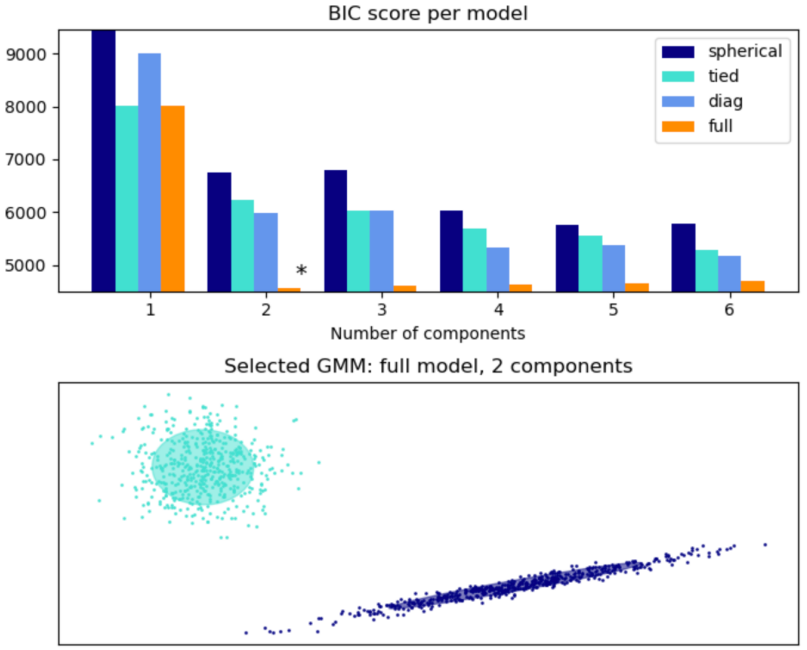
- As we increase K, the likelihood will keep increasing
- More mixtures -> more parameters -> overfits



http://staffblogs.le.ac.uk/bayeswithstata/2014/05/22/mixture-models-how-many-components/

Picking the amount of Gaussians

- Need a measure of goodness (like Elbow method in k-mean)
- Bayesian Information Criterion (BIC)
- Penalize the log likelihood from the data by the number of parameters in the model
  - -2 log L + t log (n)
  - t = number of parameters in the model
  - n = number of data points
- We want to minimize BIC



https://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_selection.html

BIC is bad use cross validation!

- Just like how I don’t recommend using elbow method for clustering
- BIC is bad use cross validation!
- Test on the goal of your model

Latent variables?

EM is all about problem formulation. You can solve the same task with different formulations.

Latent variable considerations

- Imaginary quantity meant to provide a simplified view of the process
  - GMM mixtures. Speech recognizer states. Customer segmentation.
- Real-world thing, but impossible to directly measure
  - Cause of a disease. Temperature of a star.
- Real-world thing, that is not measured because of noise/faulty sensors

Latent variables?

- Discrete latent variables: clusters/partitions data into subgroups
- Continuous latent variables: can be used for dimensionality reduction (factor analysis, etc)

EM usage examples

Image segmentation with GMM EM

- D - {r,g,b} value at each pixel
- Latent : segment where each pixel comes from
- Hyperparameters: number of mixtures (K), initial values

input

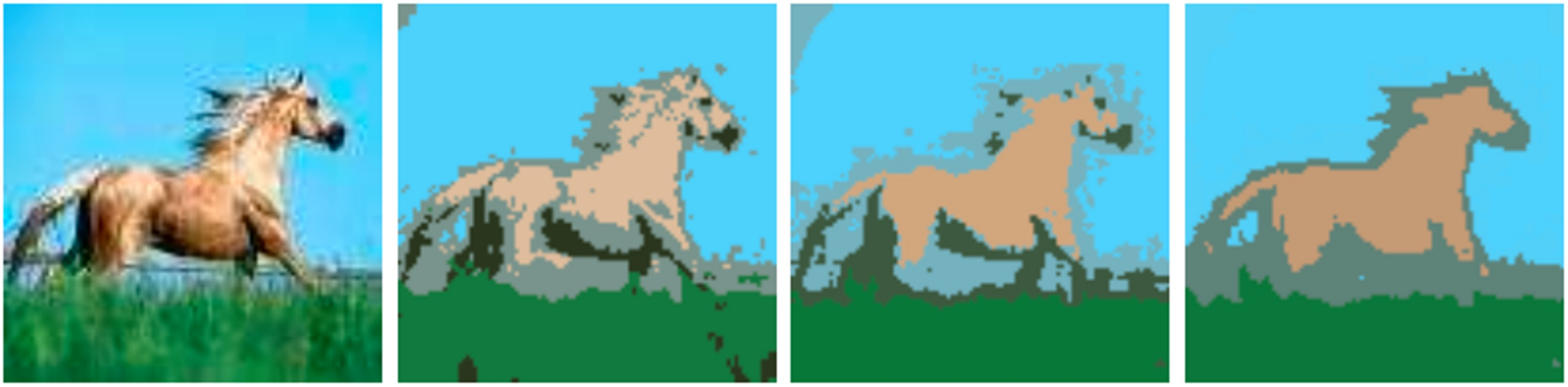


Image segmentation with GMM EM

(a) (b) (c)

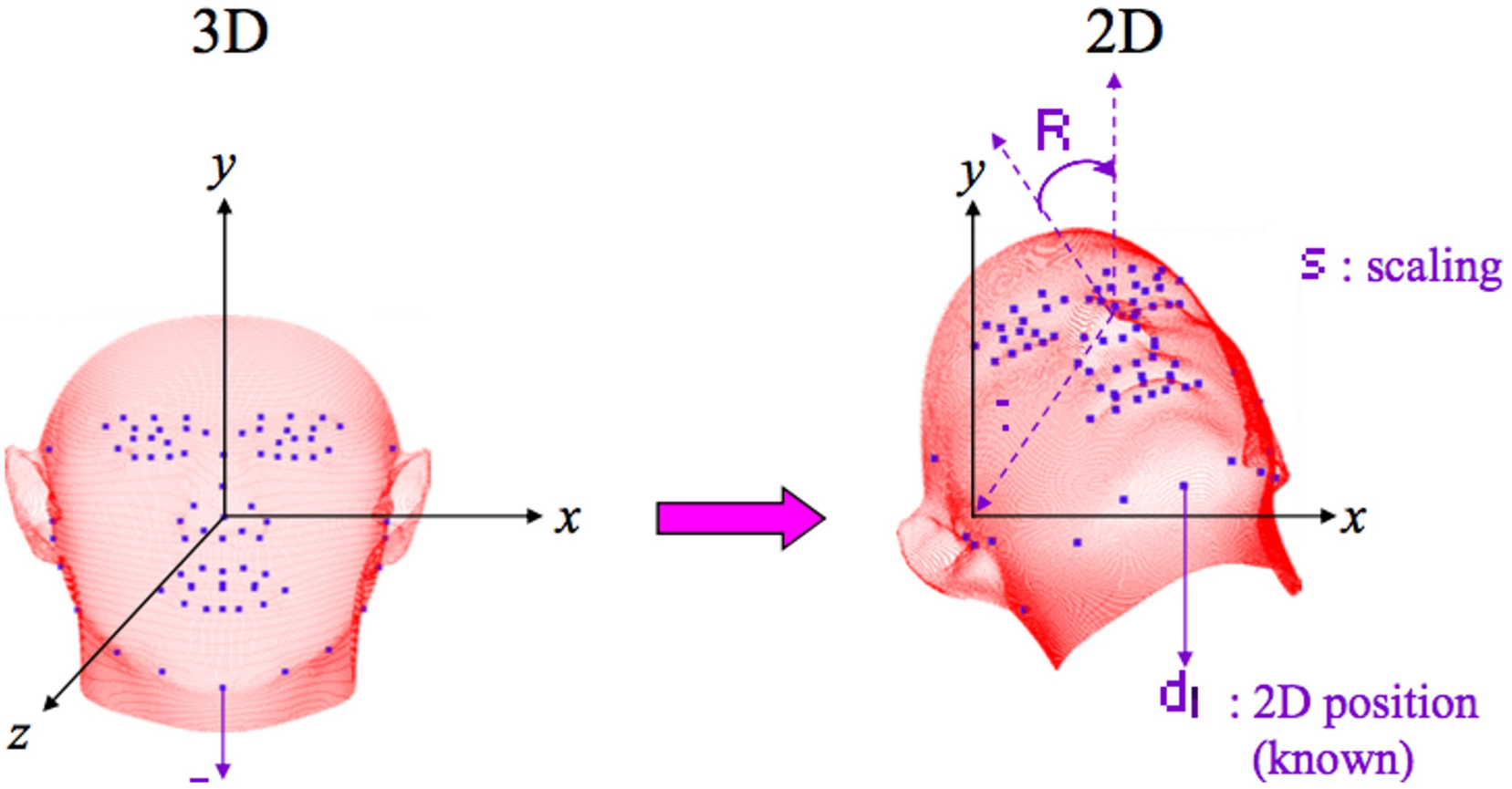
Fig. 1. Original images: (a) flower, (b) tiger, (c) bear

Fig. 2. Segmentation results ($M = 2$)

Fig. 3. Segmentation results ($M = 5$)

Zhaoxia Fu Color Image Segmentation Using Gaussian Mixture Model and EM Algorithm

Face pose estimation (estimate 3d coordinates from 2d picture)



https://www.researchgate.net/publication/2489713_Estimating_3D_Facial_Pose_using_the_EM_Algorithm

Language modeling

THE UNITED STATES CONSTITUTION

We the People of the United States, in Order to form a more perfect Union, establish Justice, insure domestic Tranquility, provide for the common defence, promote the general Welfare, and secure the Blessings of Liberty to ourselves and our Posterity, do ordain and establish this Constitution for the United States of America.

Article. I.

Section. 1.

All legislative Powers herein granted shall be vested in a Congress of the United States, which shall consist of a Senate and House of Representatives.

Section. 2.

Clause 1: The House of Representatives shall be composed of Members chosen every second Year by the People of the several States, and the Electors in each State shall have the Qualifications requisite for Electors of the most numerous Branch of the State Legislature.

Clause 2: No Person shall be a Representative who shall not have attained to the Age of twenty-five Years, and been seven Years a Citizen of the United States, and who shall not, when elected, be an Inhabitant of that State in which he shall be chosen.

Clause 3: Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians not taxed, three fifths of all other Persons. The actual

Latent variable:
Topic
P(word|topic)

For examples: see Probabilistic latent semantic analysis

MEME

Know Your Meme

Multiple EM for Motif Elicitation

From Wikipedia, the free encyclopedia

For other uses, see MEME (disambiguation).

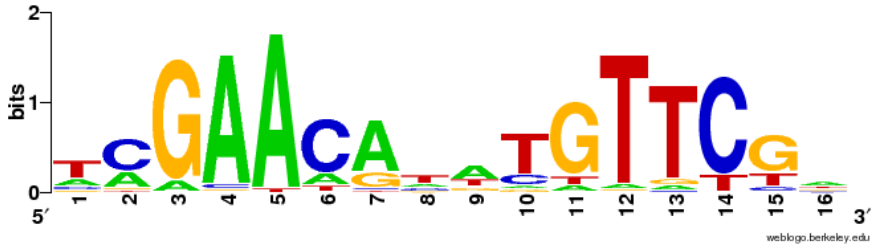
Multiple Expectation maximizations for Motif Elicitation (MEME) is a tool for discovering motifs in a group of related DNA or protein sequences.[1]

A motif is a sequence pattern that occurs repeatedly in a group of related protein or DNA sequences and is often associated with some biological function. MEME represents motifs as position-dependent letter-probability matrices which describe the probability of each possible letter at each position in the pattern. Individual MEME motifs do not contain gaps. Patterns with variable-length gaps are split by MEME into two or more separate motifs.

MEME takes as input a group of DNA or protein sequences (the training set) and outputs as many motifs as requested. It uses statistical modeling techniques to automatically choose the best width, number of occurrences, and description for each motif.

MEME is the first of a collection of tools for analyzing motifs called the MEME suite.

Contents [hide]



https://en.wikipedia.org/wiki/Multiple_EM_for_Motif_Elicitation
https://en.wikipedia.org/wiki/Position_weight_matrix

Summary

- GMM
  - Mixture of Gaussians
- EM
  - Expectation
  - Maximization

More info and exact proofs

https://seanborman.com/publications/EM_algorithm.pdf

http://cs229.stanford.edu/summer2019/cs229-notes8.pdf