Reasoning (II)

Mingsheng Long

Tsinghua University

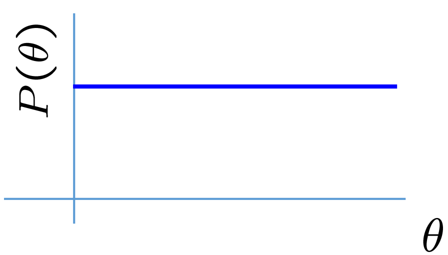
Spring 2023

Outline

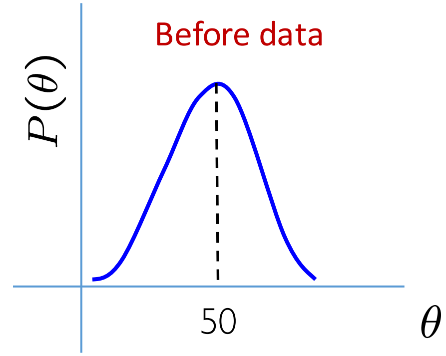
- **Probability Basics**
- Discriminative Models
- Generative Models
  - Naïve Bayes Classifier
  - Gaussian Discriminant Analysis
- Mixture Models and EM
  - Gaussian Mixture Model
  - Expectation Maximization

Bayesian Approach

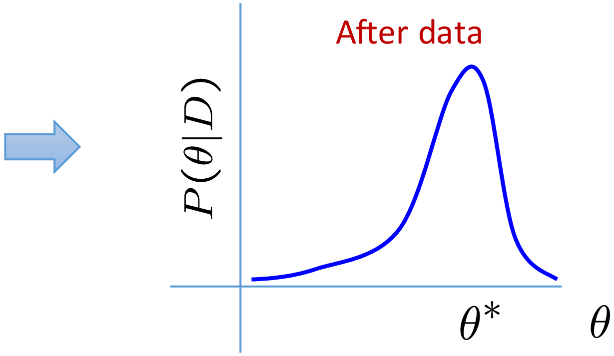
- Bayesian approaches try to reflect our belief about parameter $\boldsymbol{\theta}$.
  - In this case, we will consider $\boldsymbol{\theta}$ to be a random variable.
- Use the prior information and decide a prior distribution of $\boldsymbol{\theta}$.
- Given the data, estimate a posterior distribution over possible values of $\boldsymbol{\theta}$ with Bayes rule.



Uninformative priors: uniform distribution



We believe that $\boldsymbol{\theta}^*$ is around 50



Posterior distribution

Bayes Rule

- Bayes rule:

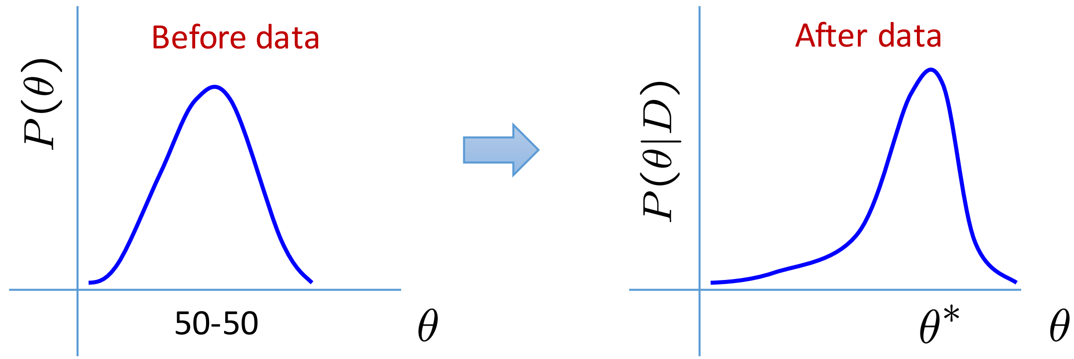
Observed Data Parameter

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- Equivalently:

$$\underbrace{p(\theta|D)}_{\text{Posterior}} \propto \underbrace{p(D|\theta)}_{\text{Likelihood}}\underbrace{p(\theta)}_{\text{Prior}}$$

Thomas Bayes



Bernoulli Distribution

- Bernoulli probability mass function (PMF):

$$P(x = 1|q) = q$$
$$P(x = 0|q) = 1 - q$$

- Mean:

$$\mathbb{E}[x] = q$$

- Multinoulli probability mass function (PMF):

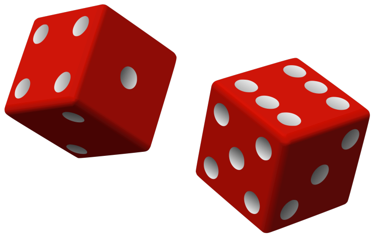
$$P(y = l|\boldsymbol{\phi}) = \phi_l$$
$$\sum_{l=1}^{C} \phi_l = 1$$

- Parameters:

$$\{\boldsymbol{\phi}\}$$

Flipping coin

Head	q
Tail	$1 - q$



Flipping dice

$$\sum_{l=1}^{6} \phi_l = 1$$

Gaussian Distribution

- Probability density function (PDF):

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$$

- Mean:

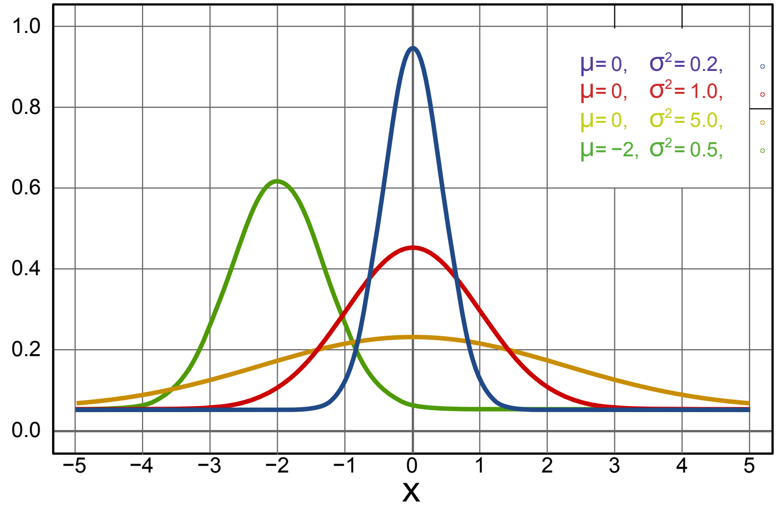
$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}$$

- Variance:

$$\mathrm{Var}[\boldsymbol{x}] = \boldsymbol{\Sigma}$$

- Parameters:

$$\{\boldsymbol{\mu},\boldsymbol{\Sigma}\}$$



Likelihood of Parametric Model

- Suppose we have a parametric model $\{p(z;\theta)|\theta \in \Theta\}$ and a sample dataset $\mathcal{D} = (z_1, \dots, z_N)$.
- The likelihood of estimated parameter $\hat{\theta} \in \Theta$ for sample $\mathcal{D}$ is

$$p(\mathcal{D};\hat{\theta}) = \prod_{n=1}^{N} p(z_n;\hat{\theta})$$

- Due to numerical instability, we prefer to work with the log-likelihood

$$\log p(\mathcal{D};\hat{\theta}) = \sum_{n=1}^{N} \log p(z_n;\hat{\theta})$$

Maximum Likelihood Estimation

- Suppose $\mathcal{D} = (z_1, \dots, z_N)$ is an i.i.d. sample from some distribution.
- Finding the maximum likelihood estimator (MLE) for parameter $\boldsymbol{\theta}$ in the parametric model $\{p(y; \theta) | \theta \in \Theta\}$ is an optimization problem:

$$\boxed{\begin{aligned}\hat{\theta} &\in \underset{\theta \in \Theta}{\operatorname{argmax}} \log p(\mathcal{D}; \theta) \\ &= \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(z_n; \theta)\end{aligned}}$$

- **Note:** MLE of a parametric model leads to a particular loss function.

MLE for Gaussian Distribution

- Recall that the density of Gaussian distribution is $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$$

- The log-density is

$$\log p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

- To estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from an i.i.d. sample $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we will maximize the log joint density

$$\sum_{i=1}^{n} \log p(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2}\log|2\pi\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})$$

MLE for Gaussian Distribution

- To estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from an i.i.d. sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we will maximize the log joint density

$$\sum_{i=1}^{n} \log p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2} \log |2\pi\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

- A solid exercise in vector and matrix differentiation. Find $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ by

$$\nabla_{\boldsymbol{\mu}} J(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0 \qquad \nabla_{\boldsymbol{\Sigma}} J(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0$$

- We get a closed-form solution:

Check: $\widehat{\boldsymbol{\Sigma}}_{\mathrm{MLE}} = \frac{n-1}{n} \boldsymbol{\Sigma}$

$$\widehat{\boldsymbol{\mu}}_{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \qquad \widehat{\boldsymbol{\Sigma}}_{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{\mathrm{MLE}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{\mathrm{MLE}})^T$$

Bayes Decision Rule

- Assumption:
  - The learning task $p(X, Y) = p(Y|X)p(X)$ can be sampled from.
- Question:
  - Given instance $\boldsymbol{x}$, how should it be classified to minimize error?
- Bayes Decision Rule:

$$\boxed{h(\boldsymbol{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}[p(Y = y | X = \boldsymbol{x})]}$$

- How to directly measure $p(Y|X)$ with a parametric model $q(Y|X, \theta)$?
  - Discriminative models!

Outline

- Probability Basics
- **Discriminative Models**
- Generative Models
  - Naïve Bayes Classifier
  - Gaussian Discriminant Analysis
- Mixture Models and EM
  - Gaussian Mixture Model
  - Expectation Maximization

Linear Regression

- In regression problem, we assume that:

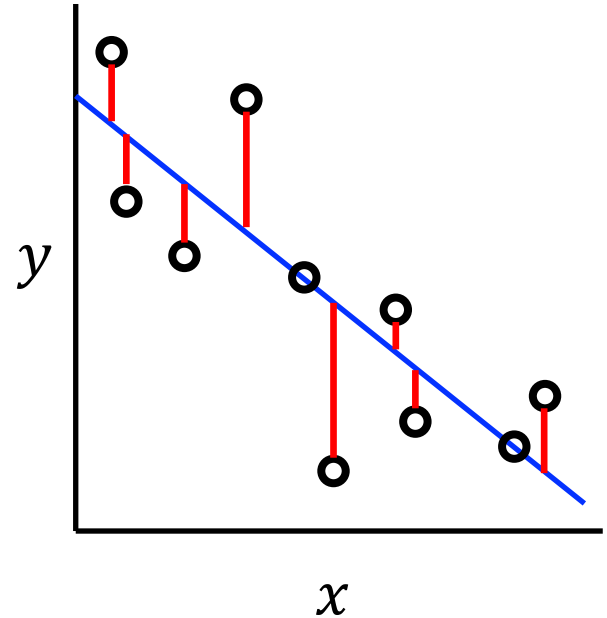
$$y \sim \mathcal{N}(\boldsymbol{w}^T \boldsymbol{x}, \sigma^2)$$

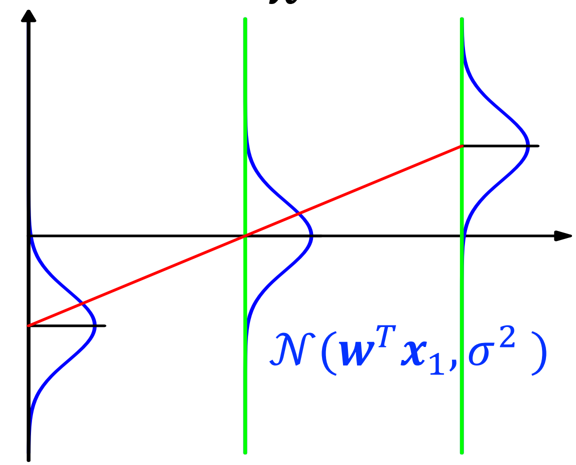
  when $\boldsymbol{y}$ is independent with each other.

- The linear regression model should give expectation of $\boldsymbol{y}$:

$$\boxed{\mathbb{E}(y|\boldsymbol{x}, \boldsymbol{w}, \sigma^2) = \boldsymbol{w}^T \boldsymbol{x}}$$

- Find the best parameter $\boldsymbol{w}$ using i.i.d. sample $\mathcal{D} = \{(x_1, y_1), \dots, (x_N, y_N)\}$.
- $\boldsymbol{\sigma}$ is useless in the final regression model.





Gaussian Linear Regression

- If we assume that $y_n|\boldsymbol{w}, \boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{w}^T\boldsymbol{x}_n, \sigma^2)$
- Then for point $(\boldsymbol{x}_n, y_n)$

$$p(y_n|\boldsymbol{w}, \boldsymbol{x}_n) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2\sigma^2}(y_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2\right\}$$

- The log-likelihood for linear regression on the whole dataset $\mathcal{D}_n$:

$$\log p(\mathcal{D}_n; \boldsymbol{w}) = \sum_{n=1}^{N} \log p(y_n|\boldsymbol{w}, \boldsymbol{x}_n)$$

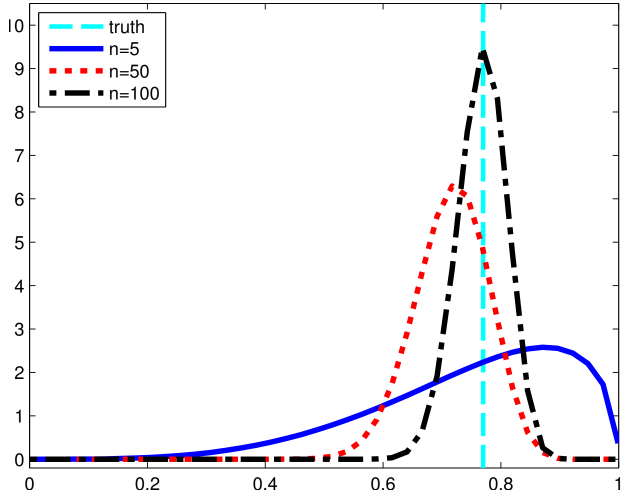
$$= \frac{N}{2}\log\frac{1}{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\boldsymbol{x}_n)^2$$

Maximum A Posteriori Estimation

$$\hat{\theta} = \arg\max_{\theta} \log p(\theta|D) = \arg\max_{\theta}\{\log p(D|\theta) + \log p(\theta)\}$$

Posterior Likelihood Prior

- MAP: Maximum a posteriori estimation of parameters $\boldsymbol{\theta}$
- We can view MLE as MAP
  - with a uniform prior distribution.
- As amount of data becomes large, posterior variance becomes small, and MAP behaves like other estimators such as MLE.



Bayesian Linear Regression

- Recall the Gaussian noise assumption for linear regression:

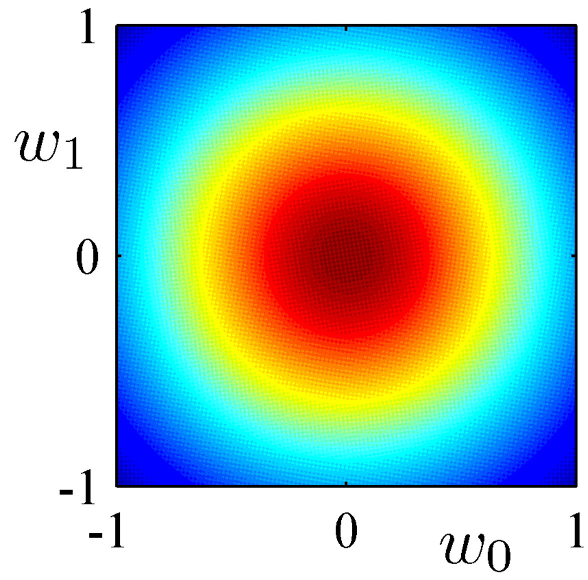
$$y_n = \boldsymbol{w}^T \boldsymbol{x}_n + \epsilon_n, \qquad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

- Then we can get:

$$y_n | \boldsymbol{w}, \boldsymbol{x}_n \sim \mathcal{N}(\boldsymbol{w}^T \boldsymbol{x}_n, \sigma^2)$$

- A common choice for the prior:

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} | 0, \alpha^{-1} \boldsymbol{I})$$



MAP and Regularization

- MAP estimation of $\boldsymbol{w}$:

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} -\log p(\boldsymbol{w}|D_n) \;= \arg\min_{\boldsymbol{w}} \{-\log p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{X}) - \log p(\boldsymbol{w})\}$$

$$-\log p(\boldsymbol{y}|\boldsymbol{w}) = -\sum_{i=1}^{n} \log p(y_i|\boldsymbol{w}, \boldsymbol{x}_i) = -\frac{n}{2}\log\frac{1}{2\pi\sigma^2} + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2$$

$$-\log p(\boldsymbol{w}) = -\frac{d}{2}\log\frac{1}{2\pi\sigma^2} + \frac{\alpha}{2}\sum_{j=1}^{d} w_j^2$$

$$\Longrightarrow \hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)^2 + \frac{\alpha}{2}\sum_{j=1}^{d} w_j^2$$

$$\Longrightarrow \hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \|\boldsymbol{X}\boldsymbol{w} - y\|_2^2 + \frac{\alpha}{\beta}\|\boldsymbol{w}\|_2^2$$

denote $\beta = \frac{1}{\sigma^2}$

Bayesian Model Averaging

- The posterior predictive distribution ($\tilde{y}$ is the prediction):

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

  - Assume there is a true model $p(y|\theta)$
  - Account for the uncertainty in θ.

- To account for model uncertainty among some models $M_1, \dots, M_h$,we use **Bayesian model averaged (BMA)** posterior predictive distribution

$$p(\tilde{y}|y) = \sum_{h=1}^{H} p(\tilde{y}|M_h, y)p(M_h|y)$$

predictive distribution under model M_h posterior model probability

Outline

- Probability Basics
- Discriminative Models
- **Generative Models**
  - **Naïve Bayes Classifier**
  - Gaussian Discriminant Analysis
- Mixture Models and EM
  - Gaussian Mixture Model
  - Expectation Maximization

Bayes Rule

- Alternative idea:
  - It is possible to switch conditioning according to Bayes rule.
  - Given any two random variables X and Y, it holds that:

$$\boxed{p(Y = y|X = \boldsymbol{x}) = \frac{p(X = \boldsymbol{x}|Y = y)p(Y = y)}{p(X = \boldsymbol{x})}}$$

- We try to model $p(X = \boldsymbol{x}|Y = y)$ and $p(Y = y)$ in this problem.
  - Generative models!
- We also write $p(Y = y|X = \boldsymbol{x}) \propto p(X = \boldsymbol{x}|Y = y)p(Y = y)$
  - $\propto$ is commonly used to avoid non-necessary normalization term.

Naïve Bayes Classifier

- Model distribution of high-dimensional data $p(X = \boldsymbol{x}|Y = y)$ is hard:
  - Because we need to find a proper description of data distribution.
  - Especially the dependency between dimensions.
- **Naïve Bayes Classifier (NBC)** assumes conditional-independence:
  - Each dimension is independent given label y:

$$p(X = \boldsymbol{x}|Y = y) = \prod_{j=1}^{d} p(x_{\cdot j}|y)$$

Naïve!

jth dimension of $\boldsymbol{x}$

  - Thus $p(y)$ and $p(x_i|y)$ can be computed directly from dataset.

Naïve Bayes Classifier

- Naïve Bayes Classifier is used when the features are all discrete.
- For binary feature, model $p(x_{\cdot j}|y)$ and $p(y)$ as **Bernoulli distribution**:

$\boxed{x_{\cdot j} \in \{0,1\}}$

$$p(x_{\cdot j} = 1|y = +1) \sim \text{Bernoulli}(\phi_j^+),$$
$$p(x_{\cdot j} = 1|y = -1) \sim \text{Bernoulli}(\phi_j^-),$$
$$p(y = +1) \sim \text{Bernoulli}(\phi)$$

- So we can estimate pareameters ϕ_j^+ and ϕ as (ϕ_j^- is similar):

$$\phi_j^+ = \frac{\sum_{i=1}^{n} \mathbf{1}\{x_{ij} = 1 \wedge y_i = +1\}}{\sum_{i=1}^{n} \mathbf{1}\{y_i = +1\}}$$

$$\phi = \frac{\sum_{i=1}^{n} \mathbf{1}\{y_i = +1\}}{n}$$

Naïve Bayes Classifier

- For feature with many values, use **multinoulli distribution** instead.

D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes

- What will Naïve Bayes Classifier predict for (Rain, Cool, High, Weak)?

$$p(\text{Rain}|\text{Yes}) = 2/4, p(\text{Cool}|\text{Yes}) = 2/4, p(\text{High}|\text{Yes}) = 2/4, p(\text{Weak}|\text{Yes}) = 3/4, p(\text{Yes}) = 4/7$$

- $p(\text{Yes}|(\text{Rain}, \text{Cool}, \text{High}, \text{Weak})) \propto 3/56$
- $p(\text{No}|(\text{Rain}, \text{Cool}, \text{High}, \text{Weak})) \propto 2/189$

Naïve Bayes Classifier

- What if $p(x_{\cdot j} = r_j | Y = +1) = \frac{\sum_{i=1}^{n} 1\{x_{\cdot j} = r_j \wedge y_i = +1\}}{\sum_{i=1}^{n} 1\{y_i = +1\}} = 0$?
  - This will cause $p(\boldsymbol{x}|Y = +1)$ to zero, no matter how large other $p(x_{\cdot l} = r_l | Y = +1)$.
  - We can add prior to solve this problem (Laplacian smoothing):

$$p(x_{\cdot j} = r_j | Y = +1) = \frac{\sum_{i=1}^{n} 1\{x_{\cdot j} = r_j \wedge y_i = +1\} + 1}{\sum_{i=1}^{n} 1\{y_i = +1\} + k_j}$$

How many values does this feature have?

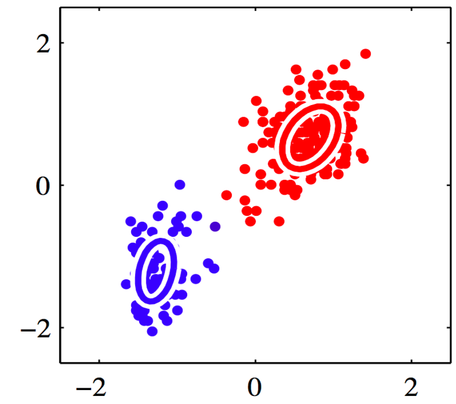
- Continuous variables:
  - We can discretize the variable.
  - Or we can use another model based on a different assumption.

Outline

- Probability Basics
- Discriminative Models
- **Generative Models**
  - Naïve Bayes Classifier
  - **Gaussian Discriminant Analysis**
- Mixture Models and EM
  - Gaussian Mixture Model
  - Expectation Maximization

Gaussian Discriminant Analysis

- Alternative methods for dataset with all continuous features:
  - Using parametric distribution to represent $p(X = \boldsymbol{x} | Y = y)$.
- A **common assumption** in classification:
  - We always assume that data points in a class is a cluster.



- So we can model $p(X = \boldsymbol{x} | Y = y)$ by **Gaussian distribution**:

$$p(X = \boldsymbol{x} | Y = +1) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_+)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_+)\right)$$

$$p(X = \boldsymbol{x} | Y = -1) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_-)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_-)\right)$$

Usually share $\boldsymbol{\Sigma}$.

Gaussian Discriminant Analysis

- We still model $p(Y=y)$ as Bernoulli distribution: $\mathrm{Bernoulli}(\phi)$.
- Now we use MLE to find the best parameter estimation:

$$\ell(\phi, \boldsymbol{\mu}_+, \boldsymbol{\mu}_-, \boldsymbol{\Sigma}) = \log \prod_{i=1}^{n} p(\boldsymbol{x}_i, y_i; \phi, \boldsymbol{\mu}_+, \boldsymbol{\mu}_-, \boldsymbol{\Sigma})$$

$$= \boxed{\log \prod_{i=1}^{n} p(\boldsymbol{x}_i | y_i; \boldsymbol{\mu}_+, \boldsymbol{\mu}_-, \boldsymbol{\Sigma})} + \boxed{\log \prod_{i=1}^{n} p(y_i | \phi)}$$

- The computing process is very similar to the process of Gaussian.
  - The main difference is that $\boldsymbol{\mu}_+, \boldsymbol{\mu}_-$ are different.

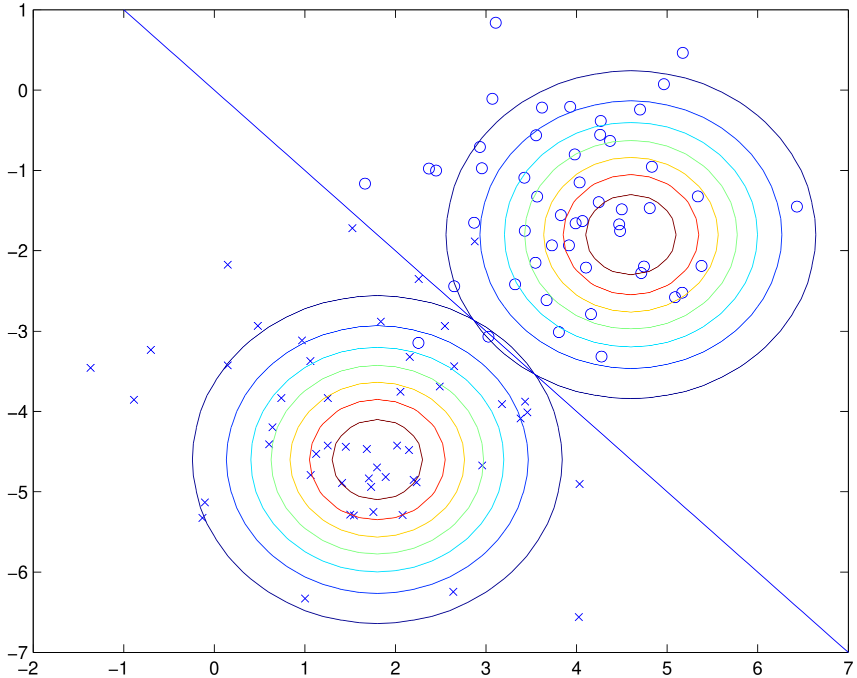
$$\phi = \frac{\sum_{i=1}^{n} \mathbf{1}\{y_i = +1\}}{n}, \boldsymbol{\mu}_+ = \frac{\sum_{i=1}^{n} \mathbf{1}\{y_i = +1\} \boldsymbol{x}_i}{\sum_{i=1}^{n} \mathbf{1}\{y_i = +1\}}, \boldsymbol{\mu}_- = \frac{\sum_{i=1}^{n} \mathbf{1}\{y_i = -1\} \boldsymbol{x}_i}{\sum_{i=1}^{n} \mathbf{1}\{y_i = -1\}}$$

$$\boldsymbol{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu}_{y_i})(\boldsymbol{x}_i - \boldsymbol{\mu}_{y_i})^T$$

Gaussian Discriminant Analysis

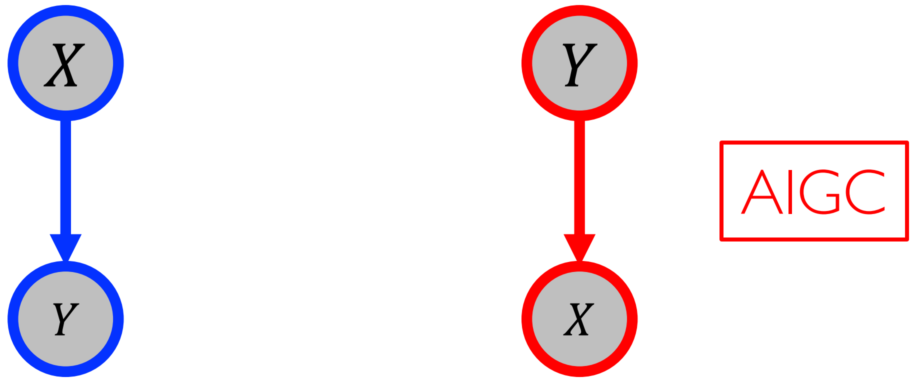
- On a test data $\boldsymbol{x}$, Gaussian Discriminant Analysis (GDA) outputs label:

$$\underset{y\in\{+1,-1\}}{\operatorname{argmax}}\,[p(\boldsymbol{x}|y)p(y)]$$



Shared $\boldsymbol{\Sigma}$ leads to "large margin model"

Discriminative vs. Generative



- Discriminative models:
  - Concentrate on the prediction of label or certain variables.
  - Usually simpler and more efficient on general data.
- Generative models:
  - Usually stronger assumption, better results on smaller data.
  - Can capture structure of data distribution and generate new data.

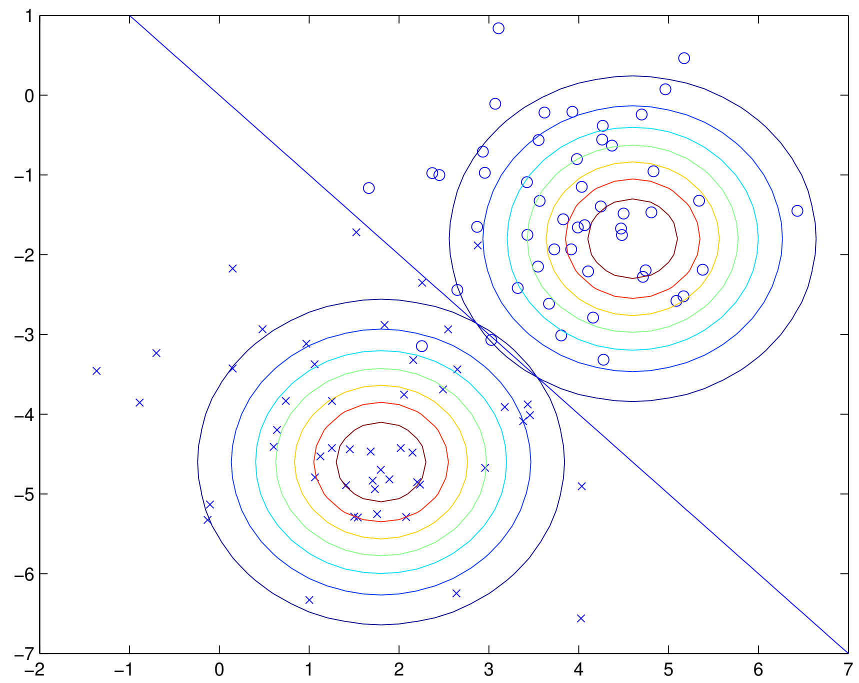
Outline

- Probability Basics
- Discriminative Models
- Generative Models
  - Naïve Bayes Classifier
  - Gaussian Discriminant Analysis
- **Mixture Models and EM**
  - **Gaussian Mixture Model**
  - Expectation Maximization

Gaussian Mixture Model

- The generating process of the GDA Model:
  - Choose $y \in \{+1, -1\}$ with $p(+1) = p(-1) = \frac{1}{2}$.
  - Choose $\boldsymbol{x}|y \sim \mathcal{N}\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_y, \boldsymbol{\Sigma}\right)$.
- We can compute $p(\boldsymbol{x})$:

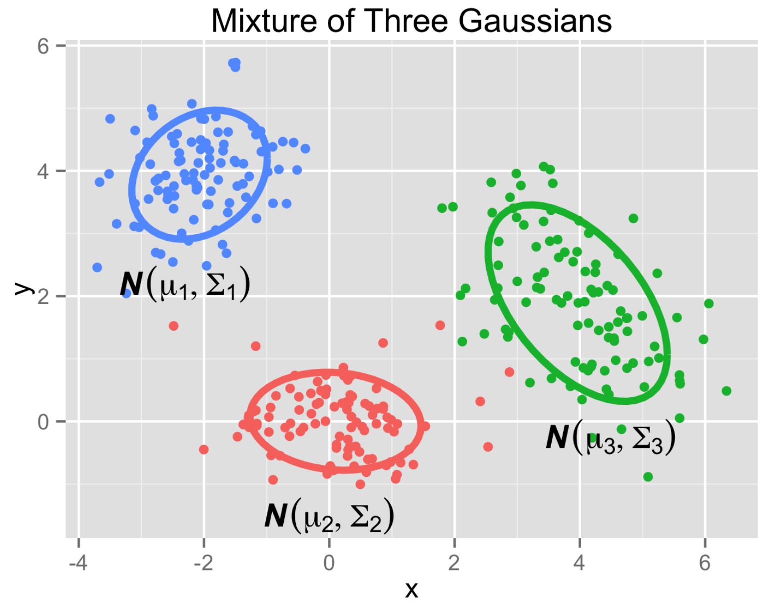
$$p(\boldsymbol{x}) = \frac{1}{2}p(\boldsymbol{x}|\boldsymbol{\mu}_{+1}, \boldsymbol{\Sigma}) + \frac{1}{2}p(\boldsymbol{x}|\boldsymbol{\mu}_{-1}, \boldsymbol{\Sigma})$$



- This is not a GDA Model, but a Gaussian Mixture Model (GMM).

Gaussian Mixture Model

- Parameters of Gaussian Mixture Model (GMM):
  - Cluster probabilities: $\pi = (\pi_1, \dots, \pi_k)$.
  - Cluster means: $\mu = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k)$.
  - Cluster covariance matrices: $\Sigma = (\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_k)$.
- Generating process of GMM:
  - First generate cluster index:
    - $z \sim (\pi_1, \dots, \pi_k)$
  - Then generate data:
    - $x \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$.
- Density: $p(\boldsymbol{x}) = \sum_{z=1}^{k} \pi_z \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$



Mixture Distribution

- A probability density $p(x)$ represents a mixture distribution
  - if we can write it as a convex combination of probability densities:

$$\boxed{p(x) = \sum_{i=1}^{k} w_i p_i(x)}$$

  - where $w_i \geq 0, \sum_{i=1}^{k} w_i = 1$, and each p_i is a probability density.
- Gaussian mixture model (GMM): $\boldsymbol{p(x)} = \sum_{z=1}^{k} \pi_z \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$.
- More constructively, let $\boldsymbol{S}$ be a set of probability distributions:
  - Choose a distribution randomly from $\boldsymbol{S}$.
  - Sample $\boldsymbol{x}$ from the chosen distribution.
    - Then $\boldsymbol{x}$ has a mixture distribution.

Generative Models for Clustering

- What do we model in unsupervised learning setting?
  - There is no longer $\boldsymbol{p(x, y)}$. We can sample from $\boldsymbol{p(x)}$ only.
- Consider a **clustering** problem.
  - Suppose there are k clusters.
  - We have a distribution assumption (Gaussian) for each cluster.
- And the whole dataset can be generated as follows:
  - Choose a random cluster $z \in \{1,2, \dots, k\}$.
  - Choose a point $\boldsymbol{x}$ from the distribution for cluster z.
- We can see that GMM is very suitable for modeling $\boldsymbol{p(x)}$.
  - **Difficulty:** We do not know which $\boldsymbol{x}$ is sampled from which z!

Gaussian Mixture Model: Learning

- Can we compute MLE of GMM directly?
- The log-likelihood for $\mathcal{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ sampled i.i.d. from a GMM is

$$\ell(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log \prod_{i=1}^{n} \sum_{z=1}^{k} \pi_z \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) = \sum_{i=1}^{n} \log \left[\sum_{z=1}^{k} \pi_z \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) \right]$$

- Plug the Gaussian density in it:

$$\ell(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \log \left[\sum_{z=1}^{k} \frac{\pi_z}{\sqrt{|2\pi \boldsymbol{\Sigma}_z|}} \exp(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_z)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_z)) \right]$$

- The sum inside the log is intractable:
  - A general challenge in mixture models, need approximate methods.

Gaussian Mixture Model: Learning

- In GDA, we know which $\boldsymbol{x}$ is sampled from which cluster z.
  - So the solution of MLE is easy to find.
  - Without z, there will be computational difficulties.
  - z is called latent variable.
- An iterative idea that solves one set of variables by fixing the others:
- Iterate between
  - Step I: Known each $(\boldsymbol{x}, z)$, find best $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.
  - Step II: Known $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, find z for each $\boldsymbol{x}$.
- We have a general method with strict theoretical foundation here!
  - Expectation-Maximization (EM)!

Outline

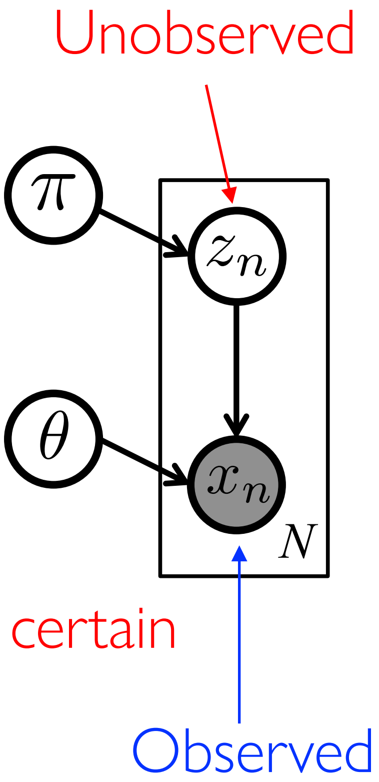
- Probability Basics
- Discriminative Models
- Generative Models
  - Naïve Bayes Classifier
  - Gaussian Discriminant Analysis
- **Mixture Models and EM**
  - Gaussian Mixture Model
  - **Expectation Maximization**

Latent Variable Model

- Two **(abstract)** sets of random variables: z and x
  - z consists of latent variables.
  - x consists of observed variables.
- Joint probability model parameterized by $\theta \in \Theta$:

$$p(x, z|\theta)$$

- A **latent variable model** is a probabilistic model for which certain variables are never observed.
  - The Gaussian mixture model is a latent variable model.
- An observation of x is called an incomplete dataset.
  - An observation of (x, z) is called a complete dataset.



Objectives

- Learning problem:
  - Given incomplete dataset $\mathcal{D} = (x_1, \dots x_n)$, find MLE

$$\hat{\theta} = \operatorname*{argmax}_{\theta} p(\mathcal{D}|\theta).$$

- Inference problem:
  - Given x, find conditional distribution over latent variable z:

$$p(z|x, \theta)$$

- Expectation-Maximization (EM) for both problems!
- For Gaussian mixture model, learning is hard, inference is easy.
- For more complicated models (next lectures), inference can be hard.

Expectation-Maximization (EM): Key Idea

- Marginal log-likelihood is **hard** to optimize:

Objective!

$$\max_{\theta} \log p(x|\theta)$$

- Typically, the complete data log-likelihood is **easy** to optimize:

$$\max_{\theta} \log p(x, z|\theta)$$

  - What if we had a distribution $q(z)$ for the latent variables z?

- Then maximize the expected complete data log-likelihood:

$$\max_{\theta} \sum_{z} q(z) \log p(x, z|\theta)$$

  - **Assumption:** EM assumes this maximization is relatively easy.

Evidence Lower Bound (ELBO)

- Let $q(z)$ be any probability function on $\mathcal{Z}$, the support of z:

$$\log p(x|\theta) = \log\left[\sum_z p(x,z|\theta)\right]$$

Objective!

$$= \log\left[\sum_z q(z)\left(\frac{p(x,z|\theta)}{q(z)}\right)\right]$$

log of an expectation

Jenson's inequality
$\log(\mathbb{E}[X]) \geq \mathbb{E}[\log(X)]$

$$\geq \boxed{\underbrace{\sum_z q(z)\log\left(\frac{p(x,z|\theta)}{q(z)}\right)}_{\mathcal{L}(q,\theta)}}$$

expectation of log

Evidence lower bound (ELBO)

MLE, EM and ELBO

- For any probability function $q(z)$, we have a lower bound on the marginal log-likelihood

$$\log p(x|\theta) \geq \mathcal{L}(q, \theta).$$

- The MLE is defined as a maximum over θ:

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} \log p(x|\theta)$$

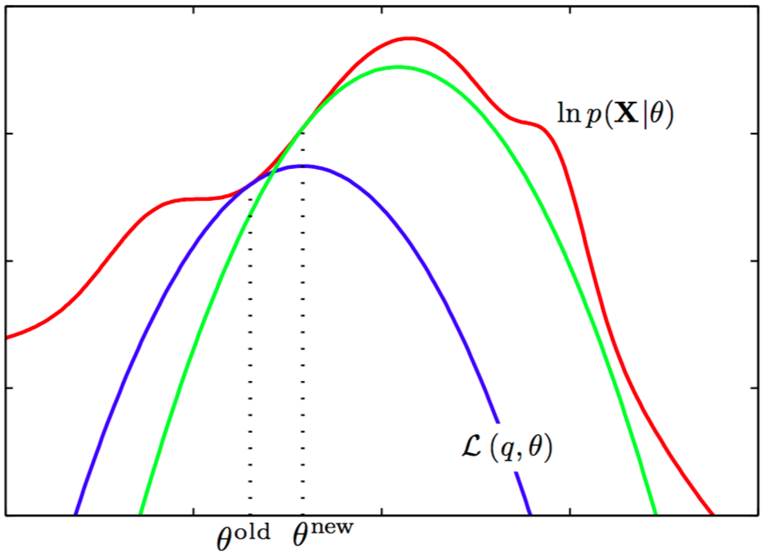
- The EM algorithm maximizes the ELBO over θ and q:

$$\hat{\theta}_{\text{EM}} = \underset{\theta}{\operatorname{argmax}} [\max_{q} \mathcal{L}(q, \theta)]$$

Lead to an **Iterative** Algorithm!

EM: Iterative Optimization

- Choose sequence of q's and θ's by **coordinate ascent.**
- EM Algorithm (high level):
  - Choose initial θ^{old}
  - Let $q^* = \underset{q}{\operatorname{argmax}}\, \mathcal{L}(q, \theta^{\text{old}})$
  - Let $\theta^{\text{new}} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(q^*, \theta)$
  - Go to Step 2, until converged.



- Will show: $p(x|\theta^{\text{new}}) \geq p(x|\theta^{\text{old}})$
  - Get sequence of θ's with monotonically increasing likelihood.
- What left: What are $\underset{q}{\operatorname{argmax}}\, \mathcal{L}(q, \theta^{\text{old}})$ and $\underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(q^*, \theta^{\text{old}})$?

ELBO via KL Divergence

- Investigate the evidence lower bound:

$$\mathcal{L}(q,\theta) = \sum_z q(z)\log\left(\frac{p(x,z|\theta)}{q(z)}\right)$$

$$= \sum_z q(z)\log\left(\frac{p(z|x,\theta)p(x|\theta)}{q(z)}\right)$$

$$= \boxed{\sum_z q(z)\log\left(\frac{p(z|x,\theta)}{q(z)}\right)} + \sum_z q(z)\log\left(p(x|\theta)\right)$$

$$= -\mathrm{KL}[q(z)||p(z|x,\theta)] + \log p(x|\theta)$$

> KL-Divergence
>
> $$\sum_z q(z)\log\left(\frac{q(z)}{p(z)}\right) = \mathrm{KL}[q(z)||p(z)]$$
>
> Properties:
>
> $\mathrm{KL}(p||q) \geq 0$,
>
> $\mathrm{KL}(p||p) = 0$.

- Amazing! We get back an equality for the marginal likelihood:

$$\log p(x|\theta) = \mathcal{L}(q,\theta) + \mathrm{KL}[q(z)||p(z|x,\theta)]$$

E-Step: Maximizing Over q for Fixed $\theta = \theta^{\text{old}}$

- Find q maximizing

$$\mathcal{L}\left(q, \theta^{\text{old}}\right) = -\text{KL}\left[q(z), p\left(z \middle| x, \theta^{\text{old}}\right)\right] + \underbrace{\log p(x|\theta^{\text{old}})}_{\text{no } q \text{ here}}$$

  - Recall $\text{KL}(p||q) \geq 0$, and $\text{KL}(p||p) = 0$

- The best q is $\boxed{q^*(z) = p(z|x, \theta^{\text{old}})}$

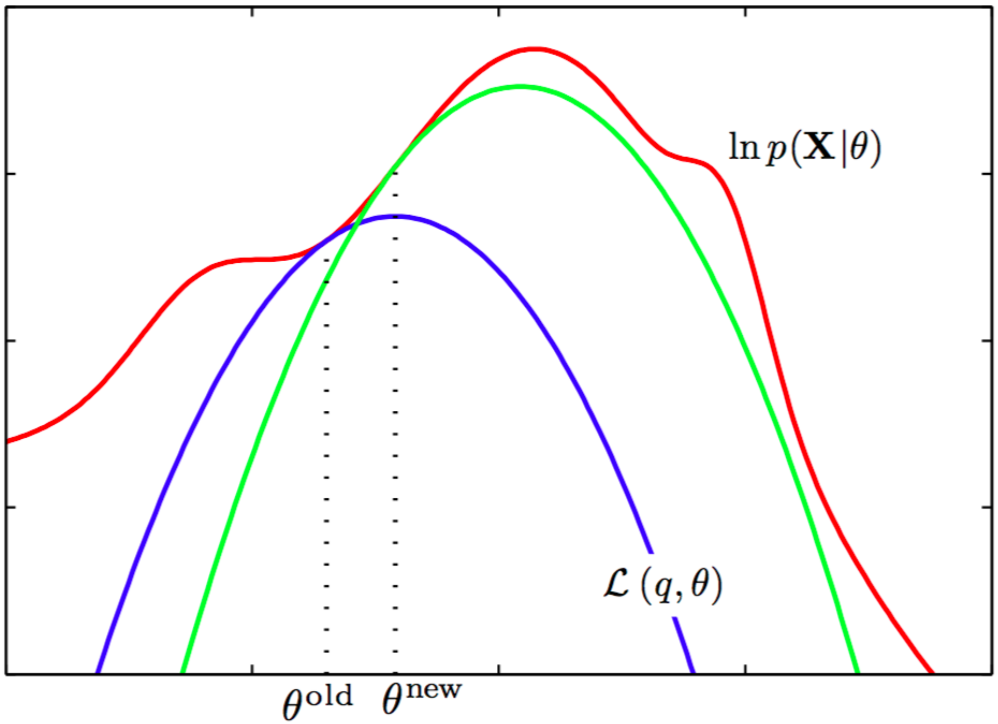
$$\mathcal{L}\left(q^*, \theta^{\text{old}}\right) = \underbrace{-\text{KL}\left[p\left(z \middle| x, \theta^{\text{old}}\right), p\left(z \middle| x, \theta^{\text{old}}\right)\right]}_{=0} + \log p(x|\theta^{\text{old}})$$

- Summary:

$$\log p\left(x \middle| \theta^{\text{old}}\right) = \mathcal{L}\left(q^*, \theta^{\text{old}}\right) \quad \text{(Tangent at } \theta^{\text{old}}\text{)}$$

$$\log p(x|\theta) \geq \mathcal{L}(q^*, \theta) \quad \forall \theta$$

Tight Lower Bound for Any Chosen θ



- For $\theta^{\mathbf{old}}$, take $q(z) = p\left(z \middle| x, \theta^{\mathbf{old}}\right)$. Then
  - $\log p(x|\theta) \geq \mathcal{L}(q, \theta)\ \ \forall \theta$. [Global lower bound]
  - $\log p\left(x \middle| \theta^{\mathbf{old}}\right) = \mathcal{L}\left(q, \theta^{\mathbf{old}}\right)$. [Lower bound is tight at $\boldsymbol{\theta}^{\mathbf{old}}$]

M-Step: Maximizing Over θ for Fixed q

- Consider maximizing the evidence lower bound (EBLO) $\mathcal{L}(q,\theta)$:

$$\mathcal{L}(q,\theta) = \sum_z q(z)\log\left(\frac{p(x,z|\theta)}{q(z)}\right)$$

$$= \underbrace{\sum_z q(z)\log\big(p(x,z|\theta)\big)}_{\mathbb{E}[\text{complete data log-likelihood}]} - \underbrace{\sum_z q(z)\log\big(q(z)\big)}_{\text{no } \boldsymbol{\theta} \text{ here}}$$

- For fixed q, maximizing $\mathcal{L}(q,\theta)$ by θ is equivalent to maximizing

$\mathbb{E}$[complete data log-likelihood]

Expectation-Maximization (EM): Algorithm

Donald Rubin

- Choose initial $\theta^{\mathbf{old}}$.
- Expectation Step
  - Let $q^*(z) = p(z|x, \theta^{\mathbf{old}})$. [$q^*$ gives best lower bound at $\theta^{\mathbf{old}}$]
  - Let $$J(\theta) \coloneqq \mathcal{L}(q^*, \theta) = \underbrace{\sum_z q^*(z) \log\left(\frac{p(x, z|\theta)}{q^*(z)}\right)}_{\text{Expectation w.r.t. } z\sim q^*(z)}$$
- Maximization Step

$$\theta^{\mathbf{new}} = \underset{\theta}{\mathrm{argmax}}\, J(\theta)$$

You can use SGD

[Equivalent to maximizing the expected complete log-likelihood.]

- Go to the Expectation Step, until converged.

EM for MAP

- Suppose we have a prior $p(\theta)$.
- Want to find MAP estimate: $\hat{\theta}_{\mathrm{MAP}} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|x)$

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

$$\log p(\theta|x) = \log p(x|\theta) + \log p(\theta) - \log p(x)$$

- Still can use our evidence lower bound on $\log p(x, \theta)$.

$$J(\theta) \coloneqq \mathcal{L}(q^*, \theta) = \sum_z q^*(z) \log\left(\frac{p(x, z|\theta)}{q^*(z)}\right)$$

- Maximization step becomes $\theta^{\mathrm{new}} = \underset{\theta}{\operatorname{argmax}}\, [J(\theta) + \log\, p(\theta)]$

GMM: E-Step

- Denote probability (responsibility) that $\boldsymbol{x}_i$ comes from cluster $\boldsymbol{j}$ by

$$\gamma_i^j = p(z = j | \boldsymbol{x} = \boldsymbol{x}_i, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boxed{q^*(z) = p(z|x, \theta^{\text{old}})}$$

  - The vector $(\gamma_i^1, \ldots, \gamma_i^k)$ is exactly the soft assignment for $\boldsymbol{x}_i$.

- From probabilistic computation:

$$\gamma_i^j = p(z = j | \boldsymbol{x}_i) = \frac{p(z = j, \boldsymbol{x}_i)}{p(\boldsymbol{x}_i)} = \frac{\pi_j \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{c=1}^k \pi_c \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}$$

- If we know $\boldsymbol{\mu}_j, \Sigma_j, \pi_j$ for all clusters $j = 1, \ldots k$, then easy to compute:

$$\gamma_i^j = \frac{\pi_j \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{c=1}^k \pi_c \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}$$

GMM: M-Step

$$\underset{\theta}{\operatorname{argmax}} \mathcal{L}(q^*, \theta) = \underset{\theta}{\operatorname{argmax}} \sum_{z} q^*(z) \log\left(\frac{p(x, z|\theta)}{q^*(z)}\right)$$

$$\boxed{\theta = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}} \qquad = \underset{\theta}{\operatorname{argmax}} \sum_{z} p(z|x, \theta^{\text{old}}) \log\left(p(x, z|\theta)\right)$$

- So we have the loss function for Gaussian Mixture Model parameters:

By MLE

$$\underset{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{j=1}^{k} \gamma_i^j \log[\pi_j \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]$$

- Let $n_c = \sum_{i=1}^{n} \gamma_i^c$ be the number of points soft-assigned to cluster c.

$$\pi_c^{\text{new}} = \frac{n_c}{n}, \boldsymbol{\mu}_c^{\text{new}} = \frac{1}{n_c} \sum_{i=1}^{n} \gamma_i^c \boldsymbol{x}_i, \boldsymbol{\Sigma}_c^{\text{new}} = \frac{1}{n_c} \sum_{i=1}^{n} \gamma_i^c (\boldsymbol{x}_i - \boldsymbol{\mu}_c^{\text{new}})(\boldsymbol{x}_i - \boldsymbol{\mu}_c^{\text{new}})^T$$

EM for GMM: Overview

- Initialize parameters $\boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\mu}$.
- **E-step.** Evaluate all responsibilities using current parameters:

$$\gamma_i^j = \frac{\pi_j \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{c=1}^{k} \pi_c \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}$$

- **M-step.** Re-estimate the parameters using the responsibilities:

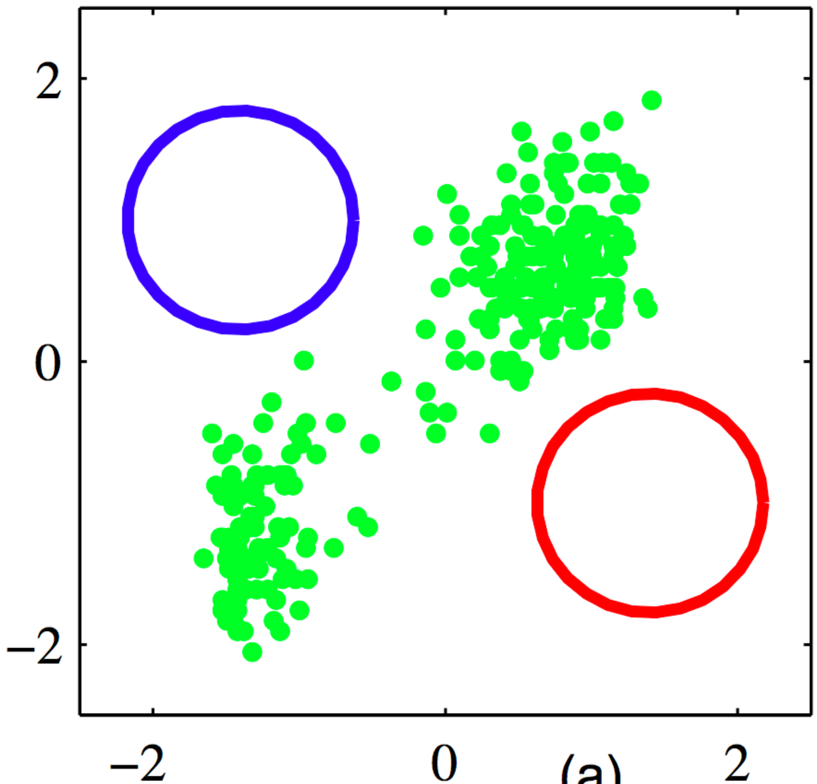
$$\pi_c^{\text{new}} = \frac{n_c}{n}$$

$$\boldsymbol{\mu}_c^{\text{new}} = \frac{1}{n_c} \sum_{i=1}^{n} \gamma_i^c \boldsymbol{x}_i, \boldsymbol{\Sigma}_c^{\text{new}} = \frac{1}{n_c} \sum_{i=1}^{n} \gamma_i^c (\boldsymbol{x}_i - \boldsymbol{\mu}_c^{\text{new}})(\boldsymbol{x}_i - \boldsymbol{\mu}_c^{\text{new}})^T$$

- Repeat E-step and M-step until log-likelihood converges.

EM for GMM

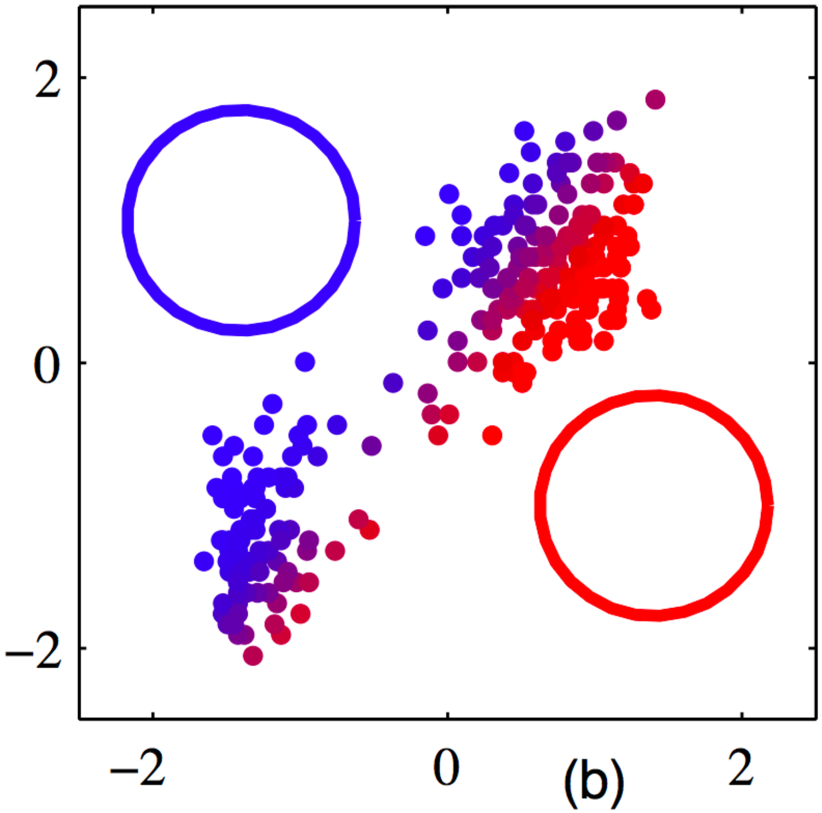
- Initialization



(a)

EM for GMM

- First soft assignment:

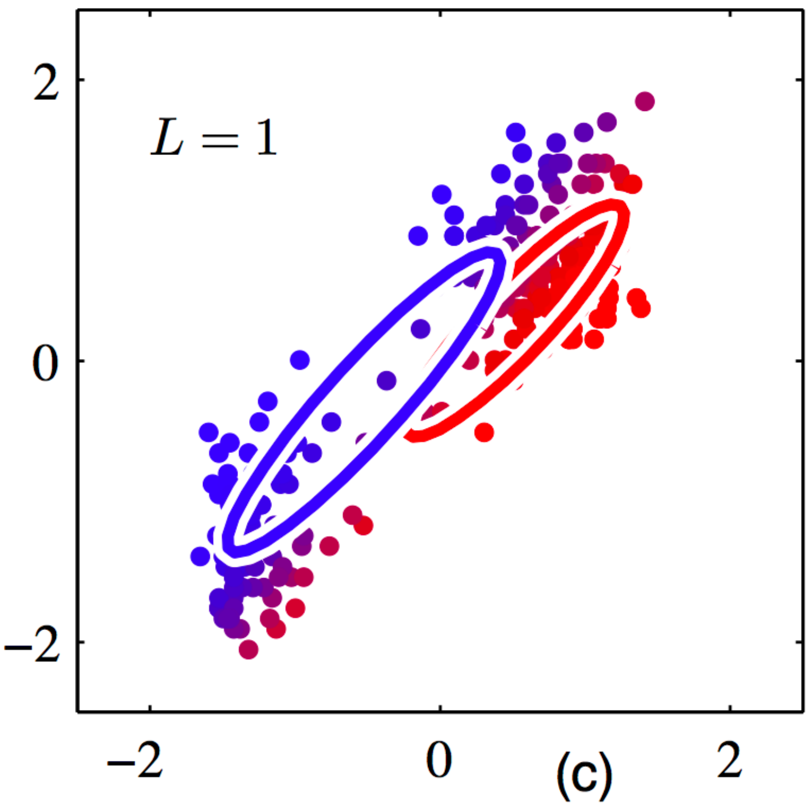


responsibilities

$$\gamma_i^j = \frac{\pi_j \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{c=1}^{k} \pi_c \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}$$

EM for GMM

- First soft assignment:



parameters

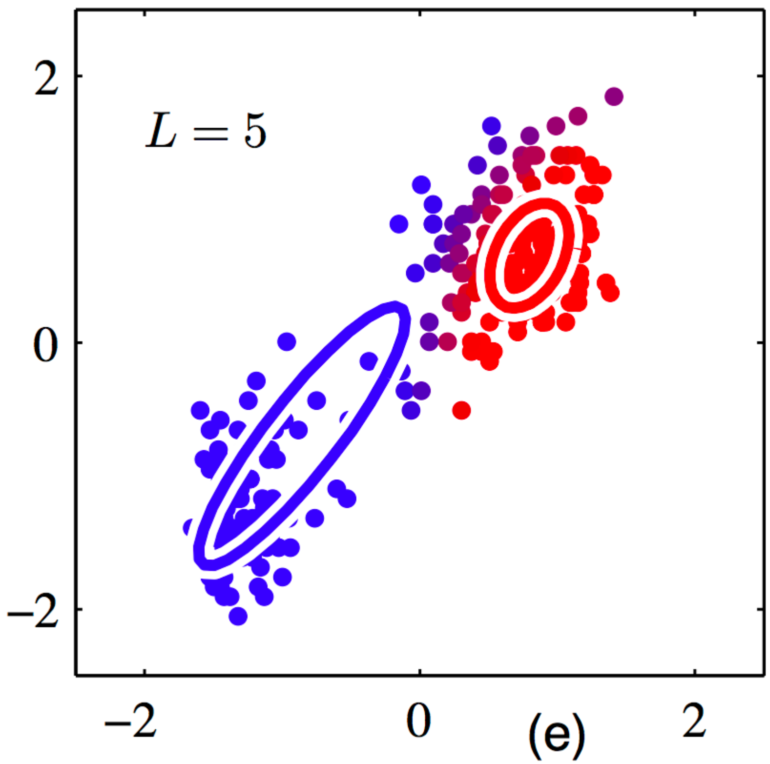
$$\pi_c^{\text{new}} = \frac{n_c}{n}$$

$$\boldsymbol{\mu}_c^{\text{new}} = \frac{1}{n_c}\sum_{i=1}^{n} \gamma_i^c \boldsymbol{x}_i$$

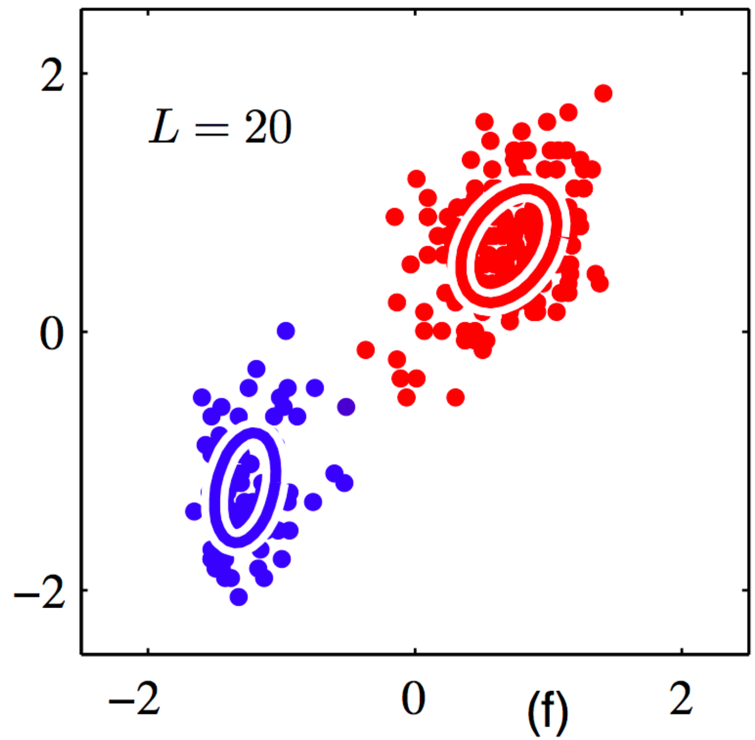
$$\boldsymbol{\Sigma}_c^{\text{new}} = \frac{1}{n_c}\sum_{i=1}^{n} \gamma_i^c (\boldsymbol{x}_i - \boldsymbol{\mu}_c^{\text{new}})(\boldsymbol{x}_i - \boldsymbol{\mu}_c^{\text{new}})^T$$

EM for GMM

- After 5 rounds of EM:



- After 20 rounds of EM:



EM and Variational Methods

- When E-step is difficult:
  - Hard to take expectation w.r.t. $q^*(z) = p(z|x, \theta^{\text{old}})$
  - For example, hierarchical latent variable models (next lectures).
- **Solution:** Restrict to distributions Q that are easy to work with.
- The evidence lower bound (ELBO) now looser:

$$q^* = \underset{q \in Q}{\operatorname{argmin}} \mathrm{KL}\left[q(z), p(z|x, \theta^{\text{old}})\right]$$

- Find an easy-to-work variational distribution q^* to approximate the inference distribution $p(z|x, \theta^{\text{old}})$.
  - This group of methods are called variational methods.

Thank You

Questions?

Mingsheng Long
mingsheng@tsinghua.edu.cn
http://ise.thss.tsinghua.edu.cn/~mlong
答疑：东主楼11区413室