1 Multiple Choice

1.1 Model space, overfitting, underfitting, regularization

- Overfitting: when a statistical model has too much flexibility, fits noise instead of signal
- Underfitting: when a statistical model is not flexible enough to capture the signal
- Regularization: additional information or constraints to reduce flexibility of the model

1.2 Bias and variance

- Tradeoff between model ability to minimize bias bias$[\hat{\theta}_n]$ and variance $\mathrm{Var}(\hat{\theta}_n)$
  - High bias, low variance $\implies$ underfitting, low training and testing accuracy
  - High variance, low bias $\implies$ overfitting, high training accuracy and low testing accuracy

$$R(f) = \mathbb{E}|Y - f(X)|^2 = (\mathbb{E}[f(X)] - Y)^2 + \mathbb{E}\left[(f(X) - \mathbb{E}[f(X)])^2\right]$$

1.3 Statistical models and unbiasedness, consistency

- **Consistent** estimator: as the sample size increases, it converges on the true parameter

$$\hat{\theta}_n \xrightarrow{P} \theta \quad \text{as} \quad n \to \infty$$

- **Unbiased** estimator: on average, it hits the true parameter

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta = 0$$

- Example: $X_1, ..., X_n \sim \mathcal{N}(\mu, 1)$

		Consistency	
		Yes	No
Unbiasedness	Yes	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\hat{\mu} = X_1$
	No	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}$	$\hat{\mu} = 6$

1.4 Naive Bayes classifier

- Naive Bayes assumption: class conditional independence

$$p_+(\mathbf{x}) = p(\mathbf{x} \mid Y = +1) = \prod_{j=1}^{d} p(x_j \mid Y = +1) \qquad p_-(\mathbf{x}) = p(\mathbf{x} \mid Y = -1) = \prod_{j=1}^{d} p(x_j \mid Y = -1)$$

1.5 Linear regression + MLE

- Linear regression: model-free approach through ordinary least squares L2 risk minimization, model-based approach through MLE
- MLE for linear regression

$$Y \mid X = \mathbf{x} \sim \mathcal{N}(\beta^T \mathbf{x}, \sigma^2)$$

$$\ell_n(\beta, \sigma^2) = \sum_{i=1}^{n} \log p_{\beta,\sigma^2}(Y_i, X_i) = \sum_{i=1}^{n} \log p_{\beta,\sigma^2}(Y_i \mid X_i) + \sum_{i=1}^{n} \log p(x_i)$$

$$
\begin{aligned}
\arg\max_{\beta,\sigma^2} \ell_n(\beta,\sigma^2) &= \arg\max_{\beta,\sigma^2} \sum_{i=1}^n \log p_{\beta,\sigma^2}(Y_i \mid X_i) \\
&= \arg\max_{\beta,\sigma^2} \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i-\beta^T X_i)^2}{2\sigma^2}} \\
&= \arg\min_{\beta,\sigma^2} \sum_{i=1}^n \frac{(Y_i-\beta^T X_i)^2}{2\sigma^2} - \log \frac{1}{\sqrt{2\pi\sigma^2}}
\end{aligned}
$$

$$
\hat{\beta}^{\text{MLE}} = \arg\min_{\beta} \sum_{i=1}^n (Y_i - \beta^T X_i)^2
$$

1.6 EM finite mixture models, K-means algorithm

- K-means algorithm is the finite mixture of K spherical Gaussian mixture models.

1.7 Classification: discriminative vs. generative

- Discriminative classification: logistic regression

$$
p(x,y) = p(y \mid x)p(x)
$$

  - Model-based logistic regression:

$$
p(y \mid x) = \frac{1}{1+e^{-yf(x)}}
$$

$$
\hat{f} = \arg\max_f \ell_n(f) = \arg\min_f \sum_{i=1}^n \log\left(1 + e^{-Y_i f(X_i)}\right)
$$

  - Model-free logistic regression:

$$
\ell^{\text{logistic}}(y, f(x)) = \log\left(1 + e^{-yf(x)}\right)
$$

$$
\hat{f} = \arg\min_f R(f) = \arg\min_f \mathbb{E}\left[\log\left(1 + e^{-Yf(X)}\right)\right]
$$

- Generative classification: QDA, LDA, Naive Bayes

$$
p(x,y) = p(x \mid y)p(y)
$$

  - Define

$$
\eta = \mathbb{P}(Y = +1)
$$

$$
p_+(x) = p(x \mid Y = +1) \qquad p_-(x) = p(x \mid Y = -1)
$$

$$
n_+ = \sum_{i=1}^n \mathbb{1}(Y_i = +1) \qquad n_- = n - n_+
$$

  - Likelihood function and $\hat{\eta}$

$$
\begin{aligned}
\ell &= \sum_{i=1}^n \log p(X_i, Y_i) = \sum_{i=1}^n \log p(X_i \mid Y_i)p(Y_i) \\
&= n_+ \log\eta + n_- \log(1-\eta) + \sum_{\substack{i=1 \\ Y_i=+1}}^n \log p_+(X_i) + \sum_{\substack{i=1 \\ Y_i=-1}}^n \log p_-(X_i)
\end{aligned}
$$

$$
\hat{\eta} = \frac{n_+}{n}
$$

- Gaussian discriminant analysis

$$X \mid Y = +1 \sim \mathcal{N}_d(\mu_+, \Sigma_+) \qquad X \mid Y = -1 \sim \mathcal{N}_d(\mu_-, \Sigma_-)$$

$$p_+(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_+|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_+)^T\Sigma_+^{-1}(x-\mu_+)} \qquad p_-(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_-|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_-)^T\Sigma_-^{-1}(x-\mu_-)}$$

$$h(x) = \begin{cases} +1 & \text{if } \frac{1}{2}r_-^2(x) - \frac{1}{2}r_+^2(x) + \frac{1}{2}\log\frac{|\Sigma_-|}{|\Sigma_+|} + \log\frac{\eta}{1-\eta} > 0 \\ -1 & \text{otherwise} \end{cases}$$

$$\hat{\mu}_+ = \frac{1}{n_+}\sum_{\substack{i=1 \\ Y_i=+1}}^{n} X_i \qquad \hat{\mu}_- = \frac{1}{n_-}\sum_{\substack{i=1 \\ Y_i=-1}}^{n} X_i$$

$$\hat{\Sigma}_+ = \frac{1}{n_+}\sum_{\substack{i=1 \\ Y_i=+1}}^{n} (X_i - \hat{\mu}_+)(X_i - \hat{\mu}_+)^T \qquad \hat{\Sigma}_- = \frac{1}{n_-}\sum_{\substack{i=1 \\ Y_i=-1}}^{n} (X_i - \hat{\mu}_-)(X_i - \hat{\mu}_-)^T$$

- Linear discriminant analysis

$$X \mid Y = +1 \sim \mathcal{N}_d(\mu_+, \Sigma) \qquad X \mid Y = -1 \sim \mathcal{N}_d(\mu_-, \Sigma)$$

$$h(x) = \begin{cases} +1 & \text{if } (\mu_+ - \mu_-)^T\Sigma^{-1}x + \frac{1}{2}\mu_-^T\Sigma^{-1}\mu_- - \frac{1}{2}\mu_+^T\Sigma^{-1}\mu_+ + \log\frac{\eta}{1-\eta} > 0 \\ & \iff \beta^T x + \beta_0 > 0 \\ -1 & \text{otherwise} \end{cases}$$

$$\hat{\Sigma} = \frac{n_+\hat{\Sigma}_+ + n_-\hat{\Sigma}_-}{n}$$

- Diagonal LDA

$$X_j \mid Y = +1 \sim \mathcal{N}\left(\mu_{+j}, \sigma_j^2\right) \qquad X_j \mid Y = -1 \sim \mathcal{N}\left(\mu_{-j}, \sigma_j^2\right)$$

$$p(x \mid Y = +1) = \sum_{j=1}^{d} p(x_j \mid Y = +1) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_j-\mu_{+j})^2}{2\sigma_j^2}}$$

- LDA and LLR have the same model space since they have $f(x) = \beta^T x + \beta_0$. But they are not the same model – LLR imposes no constraints on $p(x)$, while LDA requires $p(x)$ to be a mixture of two Gaussians.

2 Short Answer

2.1 Modes of stochastic convergence

Convergence in probability	**Convergence in distribution**
$\lim_{n\to\infty} \mathbb{P}(\lvert X_n - X\rvert > \epsilon) = 0$	$\lim_{n\to\infty} F_n(x) = F(x)$
for all $\epsilon > 0$	for x where F is continuous

Law of large numbers	**Central limit theorem**
$\bar{X}_n \xrightarrow{P} \mu$	$\bar{X}_n \xrightarrow{D} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ $S_n \xrightarrow{D} \mathcal{N}\left(n\mu, n\sigma^2\right)$ $\sqrt{n}\left(\frac{\bar{X}_n-\mu}{\sigma}\right) \xrightarrow{D} \mathcal{N}(0,1)$

- CLT $\implies$ LLN

Proof. From CLT, we know

$$\bar{X}_n - \mu \xrightarrow{D} \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$$

Then by Chebyshev's inequality,

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty$$

$\square$

- $X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$
- $X_n \xrightarrow{D} c \implies X_n \xrightarrow{P} c$ for constant c.

Proof. Suppose $X_n \xrightarrow{D} c$. From the definition of convergence in distribution and the cdf of a constant, we have

$$\lim_{n\to\infty} F_n(x) = \begin{cases} 1 & x \geq c \\ 0 & x < c \end{cases} \implies \begin{cases} \lim_{n\to\infty} F_n\left(c+\epsilon\right) = 1 \\ \lim_{n\to\infty} F_n\left(c - \frac{\epsilon}{2}\right) = 0 \end{cases}$$

In particular, this means for any $\epsilon > 0$, Thus

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(|X_n - c| > \epsilon) &= \lim_{n\to\infty} \mathbb{P}(X_n - c > \epsilon) + \mathbb{P}(X_n - c < -\epsilon) \\
&= \lim_{n\to\infty} \mathbb{P}(X_n > \epsilon + c) + \mathbb{P}(X_n < c - \epsilon) \\
&\leq \lim_{n\to\infty} \mathbb{P}(X_n > \epsilon + c) + \mathbb{P}\left(X_n \leq c - \frac{\epsilon}{2}\right) \\
&= \lim_{n\to\infty} \mathbb{P}(X_n > \epsilon + c) && \text{from (1)} \\
&= 1 - \lim_{n\to\infty} \mathbb{P}(X_n \leq \epsilon + c) \\
&= 0 && \text{from (2)}
\end{aligned}$$

$$\implies \lim_{n\to\infty} \mathbb{P}(|X_n - c| > \epsilon) = 0$$

$\square$

2.2 Lasso vs OLS

- OLS and Lasso estimators

$$\hat{\beta}^{\text{OLS}} = \arg\min_{\beta} \|Y - \mathbb{X}\beta\|_2^2 = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$$

$$\hat{\beta}^{\text{lasso}} = \begin{cases} \hat{\beta}^{\lambda} = \arg\min_{\beta} \|Y - \mathbb{X}\beta\|_2^2 + \lambda \|\beta\|_1 \\ \hat{\beta}^{t} = \arg\min_{\substack{\beta \\ \|\beta\|_2^2 \leq t}} \|Y - \mathbb{X}\beta\|_2^2 \end{cases}$$

Lasso estimator is useful for high-dimensional data, where $\mathbb{X}^T\mathbb{X}$ is not invertible.

2.3 Naive Bayes classification

- Under the Naive Bayes assumption,

$$\log \frac{p(Y=+1 \mid X=\mathbf{x})}{p(Y=-1 \mid X=\mathbf{x})} = \log \frac{p_+(\mathbf{x})\eta}{p_-(\mathbf{x})(1-\eta)} = \sum_{j=1}^{d} \log \frac{p(x_j \mid Y=+1)}{p(x_j \mid Y=-1)} + \log \frac{\eta}{1-\eta}$$

$$\#\text{parameters} = (|X_1|-1)\cdot|Y| + \cdots + (|X_d|-1)\cdot|Y| + (|Y|-1)$$

- Without the Naive Bayes assumption,

$$\log \frac{p(Y=+1 \mid X=\mathbf{x})}{p(Y=-1 \mid X=\mathbf{x})} = \log \frac{p_+(\mathbf{x})\eta}{p_-(\mathbf{x})(1-\eta)} = \log \frac{p(\mathbf{x} \mid Y=+1)}{p(\mathbf{x} \mid Y=-1)} + \log \frac{\eta}{1-\eta}$$

$$\#\text{parameters} = (|X_1|\cdots|X_d|-1)\,|Y| + (|Y|-1)$$

2.4 Lasso, Ridge, OLS regularization

- Elastic-net estimator

$$\hat{\beta}^{\text{elastic}} = \arg\min_{\beta} \|Y - \mathbb{X}\beta\|_2^2 + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2) \qquad \begin{cases} \alpha = 1 \implies \text{Lasso} \\ \alpha = 0 \implies \text{Ridge} \end{cases}$$

- Steps for choosing hyperparameters λ and α
  1. (1) Using $\alpha = 1$, fit a Lasso regularization path plot.
  2. (2) Using $\alpha = 0.6$, fit a new regularization path plot. Observe whether there is a significant change.
     - No $\implies$ use $\alpha = 1$ (Lasso) for sparsity advantage
     - Yes $\implies$ use $\alpha = 0.6$ (elastic net) because $\alpha = 1$ is unstable
  3. (3) Given α, pick λ through cross-validation.

- Regularization path plot: For given α, consider a pool of tuning parameters $\Lambda = \{\lambda_1, ..., \lambda_K\}$, and the corresponding regression estimators $\hat{\beta}^{\lambda_1}, ..., \hat{\beta}^{\lambda_K}$. The regularization path plot represents the evolution of β with λ.

- J-fold cross validation: Partition data $\mathcal{D}$ into J equal-sized parts, $\mathcal{D}_1, ..., \mathcal{D}_J$. For each of $\lambda_1, ..., \lambda_K$, calculate

$$CV(k) = \frac{1}{J}\sum_{j=1}^{J} DS_j(k) \quad \text{where} \quad DS_j(k) = \frac{1}{|\mathcal{D}_j|}\sum_{i\in\mathcal{D}_j}\left(Y_i - \left(\hat{\beta}^{\lambda_k}\right)^T X_i\right)^2$$

  Finally, pick the λ_k with the smallest $CV(k)$.

3 Classification, Bayes Rule, Bayes Risk, Bayes Formula

- Bayes rule

$$h(x) = \begin{cases} 1 & \text{if } \mathbb{P}(Y=1 \mid X=x) > \mathbb{P}(Y=0 \mid X=x) \\ 0 & \text{otherwise} \end{cases} \qquad (\star)$$

  Other ways to write $(\star)$:

$$\mathbb{P}(Y=1 \mid X=x) > \frac{1}{2} \qquad \log\frac{\mathbb{P}(Y=+1 \mid X=x)}{\mathbb{P}(Y=-1 \mid X=x)} > 0 \qquad \log\frac{p_+(x)}{p_-(x)} + \log\frac{\eta}{1-\eta} > 0$$

- Bayes risk: $R(h) = \mathbb{P}(Y \neq h(X))$

- Bayes theorem

$$\mathbb{P}(Y=y \mid X=x)=\frac{\mathbb{P}(X=x \mid Y=y)\mathbb{P}(Y=y)}{\sum\limits_y \mathbb{P}(X=x \mid Y=y)\mathbb{P}(Y=y)}$$

- Example (of all three): Suppose that $Y \in \{0,1\}$, $\mathbb{P}(Y=1)=\frac{1}{2}$, and the distribution of $X \mid Y$ is specified by

$$\mathbb{P}(X=x \mid Y=0)=\begin{cases}\frac{1}{3} & x=1\\ \frac{2}{3} & x=2\end{cases} \qquad \mathbb{P}(X=x \mid Y=1)=\begin{cases}\frac{1}{3} & x=2\\ \frac{2}{3} & x=3\end{cases}$$

From law of total probability,

$$\mathbb{P}(X=x)=\mathbb{P}(X=x \mid Y=0)\mathbb{P}(Y=0)+\mathbb{P}(X=x \mid Y=1)\mathbb{P}(Y=1)=\begin{cases}\frac{1}{3}\cdot\frac{1}{2}+0\cdot\frac{1}{2}=\frac{1}{6} & x=1\\ \frac{2}{3}\cdot\frac{1}{2}+\frac{1}{3}\cdot\frac{1}{2}=\frac{1}{2} & x=2\\ \frac{2}{3}\cdot\frac{1}{2}+0\cdot\frac{1}{2}=\frac{1}{3} & x=3\end{cases}$$

From Bayes theorem,

$$\mathbb{P}(Y=1 \mid X=x)=\frac{\mathbb{P}(X=x \mid Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(X=x)}=\begin{cases}\frac{0\cdot\frac{1}{2}}{\frac{1}{6}} & x=1\\ \frac{\frac{1}{3}\cdot\frac{1}{2}}{\frac{1}{2}} & x=2\\ \frac{\frac{2}{3}\cdot\frac{1}{2}}{\frac{1}{3}} & x=3\end{cases} = \begin{cases}0 & x=1\\ \frac{1}{3} & x=2\\ 1 & x=3\end{cases}$$

Therefore Bayes rule is given by

$$h(x)=\begin{cases}1 & \text{if } x=3\\ 0 & \text{if } x=1,2\end{cases}$$

Bayes risk is given by

$$\begin{aligned}\mathbb{P}(Y\neq h(X)) &= \mathbb{E}_X[\mathbb{P}(Y\neq h(X) \mid X)]\\ &= \sum_{x=1}^{3}\mathbb{P}(Y\neq h(x) \mid X=x)\mathbb{P}(X=x)\\ &= \mathbb{P}(Y\neq 0 \mid X=1)\cdot\frac{1}{6}+\mathbb{P}(Y\neq 0 \mid X=2)\cdot\frac{1}{2}+\mathbb{P}(Y\neq 1 \mid X=3)\cdot\frac{1}{3}\\ &= \mathbb{P}(Y=1 \mid X=1)\cdot\frac{1}{6}+\mathbb{P}(Y=1 \mid X=2)\cdot\frac{1}{2}+\mathbb{P}(Y=0 \mid X=3)\cdot\frac{1}{3}\\ &= 0\cdot\frac{1}{6}+\frac{1}{3}\cdot\frac{1}{2}+0\cdot\frac{1}{3}=\frac{1}{6}\end{aligned}$$

4 EM algorithm for latent variable models

- Setup

$$\ell(\psi)=\sum_{i=1}^{n}\log p_\psi(X_i) \qquad \begin{cases}F_{\psi^{(t)}}(\psi)\le \ell(\psi)\\ F_{\psi^{(t)}}=\ell\left(\psi^{(t)}\right)\end{cases}$$

- EM Algorithm: Initialize $\psi^{(0)}$

For $t=0,1,2,\ldots$ (until convergence)

$$\begin{cases}\text{E-step: construct } F_{\psi^{(t)}}(\psi)=\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{k}\gamma_{ij}^{(t+1)}\log\frac{p_\psi(X_i,Z_i=j)}{\gamma_{ij}^{(t+1)}}\\ \qquad\qquad \text{or simply } \gamma_{ij}^{(t+1)}=p_{\psi^{(t)}}(Z_i=j \mid X_i)\\ \text{M-step: } \psi^{(t+1)}=\arg\max\limits_{\psi}F_{\psi^{(t)}}(\psi)\end{cases}$$

- EM algorithm for finite mixture models

$$p_\psi(x) = \sum_{j=1}^{k} p_{\theta_j}(x)\eta_j \qquad \text{where } \eta_j \geq 0 \text{ and } \sum_{j=1}^{k} \eta_j = 1$$

Initialize $\psi^{(0)} = \left(\left\{ \eta_j^{(0)} \right\}_{j=1}^{k}, \left\{ \theta_j^{(0)} \right\}_{j=1}^{k} \right)$

E-step:

$$\gamma_{ij}^{(t+1)} = p_\psi(Z_i = j \mid X_i) = \frac{p_{\theta_j^{(t)}}(X_i)\eta_j^{(t)}}{\sum\limits_{\ell=1}^{k} p_{\theta_\ell^{(t)}}(X_i)\eta_\ell^{(t)}}$$

M-step:

$$F_{\psi^{(t)}}(\psi) = \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij}^{(t+1)} \log \frac{p_\psi(Z_i = j \mid X_i) p_\psi(X_i)}{\gamma_{ij}^{(t+1)}} = \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij}^{(t+1)} \log p_\psi(X_i)$$

$$\begin{aligned}
\eta_j^{(t+1)} \leftarrow \arg\max_{\theta_j} \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij}^{(t+1)} \log p_\psi(X_i) &= \arg\max_{\theta_j} \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij}^{(t+1)} \log \sum_{j=1}^{k} p_{\theta_j}(x)\eta_j \\
&= \arg\max_{\theta_j} \sum_{i=1}^{n} \gamma_{ij}^{(t+1)} \log \eta_j
\end{aligned}$$

$$\frac{\delta}{\delta \eta_j} F_{\theta^{(t)}}(\theta) = \frac{\delta}{\delta \eta_j} \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij}^{(t+1)} \log(\eta_j)$$

$$\mathcal{L} = \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij}^{(t+1)} \log(\eta_j) - \lambda \left(\sum_{j=1}^{k} \eta_j - 1 \right)$$

$$\frac{\delta \mathcal{L}}{\delta \eta_j} = \frac{\sum\limits_{i=1}^{n} \gamma_{ij}^{(t+1)}}{\eta_j} - \lambda = 0$$

$$\eta_j = \frac{\sum\limits_{i=1}^{n} \gamma_{ij}^{(t+1)}}{\lambda}$$

$$1 = \sum_{j=1}^{k} \eta_j = \frac{\sum\limits_{j=1}^{k}\sum\limits_{i=1}^{n} \gamma_{ij}^{(t+1)}}{\lambda} = \frac{\sum_{i=1}^{n} 1}{\lambda} = \frac{n}{\lambda} \implies \lambda = n$$

$$\eta_j^{(t+1)} = \frac{\sum\limits_{i=1}^{n} \gamma_{ij}^{(t+1)}}{n}$$

$$\theta_j^{(t+1)} \leftarrow \arg\max_{\theta_j} \sum_{i=1}^{n}\sum_{j=1}^{k} \gamma_{ij}^{(t+1)} \log \sum_{j=1}^{k} p_{\theta_j}(x)\eta_j = \arg\max_{\theta_j} \sum_{i=1}^{n} \gamma_{ij}^{(t+1)} \log p_{\theta_j}(x)$$

- EM algorithm for mixture of K Gaussians

$$Z \sim \text{Multi}(\eta_1, ..., \eta_k)$$

$$X_i \mid Z_i = j \sim \mathcal{N}(\mu_j, \Sigma_j)$$

Initialize $\theta^{(0)} = \left(\left\{\eta_j^{(0)}\right\}_{j=1}^k, \left\{\mu_j^{(0)}\right\}_{j=1}^k, \left\{\Sigma_j^{(0)}\right\}_{j=1}^k\right)$

E-step:

$$\gamma_{ij}^{(t+1)} = \frac{p_{\mu_j^{(t)},\Sigma_j^{(t)}}(X_i)\eta_j^{(t)}}{\sum_{\ell=1}^k p_{\mu_\ell^{(t)},\Sigma_\ell^{(t)}}(X_i)\eta_\ell^{(t)}}$$

M-step:

$$\begin{aligned} F_{\theta^{(t)}}(\theta) &= \sum_{i=1}^n \sum_{j=1}^k \gamma_{ij}^{(t+1)} \log\left(\frac{p_\theta(X_i \mid Z_i = j)\eta_j}{\gamma_{ij}^{(t+1)}}\right) \\ &= \sum_{i=1}^n \sum_{j=1}^k \left(\log(p_\theta(X_i \mid Z_i = j)) + \log\eta_j - \log\gamma_{ij}^{(t+1)}\right) \end{aligned}$$

$$p_\theta(X_i \mid Z_i = j) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_j|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_i-\mu_j)^T\Sigma_j^{-1}(x_i-\mu_j)}$$

$$\log(p_\theta(X_i \mid Z_i = j)) = -\frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma_j^{-1}| - \frac{1}{2}(X_i-\mu_j)^T\Sigma_j^{-1}(X_i-\mu_j)$$

$$F_{\theta^{(t)}}(\theta) = \sum_{i=1}^n \sum_{j=1}^k \gamma_{ij}^{(t+1)}\left(-\frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma_j^{-1}| - \frac{1}{2}(X_i-\mu_j)^T\Sigma_j^{-1}(X_i-\mu_j) + \log\eta_j - \log\gamma_{ij}^{(t+1)}\right)$$

$$\begin{aligned} \frac{\delta}{\delta\mu_j} F_{\theta^{(t)}}(\theta) &= \frac{\delta}{\delta\mu_j} \sum_{i=1}^n \sum_{j=1}^k -\frac{1}{2}\gamma_{ij}^{(t+1)}(X_i-\mu_j)^T\Sigma_j^{-1}(X_i-\mu_j) \\ &= -\sum_{i=1}^n \gamma_{ij}^{(t+1)}(X_i-\mu_j) \quad \stackrel{\text{SET}}{=} \quad 0 \end{aligned}$$

$$\mu_j^{(t+1)} \leftarrow \frac{\sum\limits_{i=1}^n \gamma_{ij}X_i}{\sum\limits_{i=1}^n \gamma_{ij}}$$

$$\Sigma_j^{(t+1)} \leftarrow \arg\max_{\Sigma_j} \sum_{i=1}^n \sum_{j=1}^k \frac{1}{2}\gamma_{ij}^{(t+1)}\left(\log|\Sigma_j^{-1}| - (X_i-\mu_j)^T\Sigma_j^{-1}(X_i-\mu_j)\right)$$

$$\frac{\delta}{\delta\Sigma_j^{-1}}\log|\Sigma_j^{-1}| = \Sigma$$

$$\frac{\delta}{\delta\Sigma_j^{-1}}(X_i-\mu_j)^T\Sigma_j^{-1}(X_i-\mu_j) = -(X_i-\mu_j)^T(X_i-\mu_j)$$

$$\frac{\delta}{\delta\Sigma_j^{-1}} F_{\theta^{(t)}}(\theta) = \frac{1}{2}\sum_{i=1}^n \gamma_{ij}^{(t+1)}\left(\Sigma_j - (X_i-\mu_j)^T(X_i-\mu_j)\right) \quad \stackrel{\text{SET}}{=} \quad 0$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^n \gamma_{ij}^{(t+1)}\left((X_i-\mu_j)^T(X_i-\mu_j)\right)}{\sum_{i=1}^n \gamma_{ij}^{(t+1)}}$$

5 K means algorithm

- Theorem: If $\ell(\psi) \leq C$ for all ψ, then the EM algorithm converges to a local maximum.

  Proof. For all t,

  $$\ell\left(\psi^{(t)}\right) = F_{\psi^{(t)}}\left(\psi^{(t)}\right) \leq F_{\psi^{(t)}}\left(\psi^{(t+1)}\right) \leq \ell\left(\psi^{(t+1)}\right)$$

  Thus $\ell\left(\psi^{(t)}\right) \leq \ell\left(\psi^{(t+1)}\right) \leq \cdots$ is a non-decreasing sequence that has an upper-bound C. Therefore the sequence converges. $\square$

- Convergence of K-means: Since K-means is an EM algorithm, we know that the likelihood of the cluster centers can only increase at each step. Now, if we have to partition n data points into K clusters, then there are exactly $\binom{n}{K}$ possible clusterings. Since we have a finite, non-decreasing sequence, it must converge.

- EM algorithm for K means

  E-step:

  $$\gamma_{ij} = p(Z_i = j \mid X_i) = \frac{p(X_i \mid Z_i = j)\eta_j}{\sum_{\ell=1}^{k} p(X_i \mid Z_i = \ell)\eta_\ell} = \frac{\frac{\eta_j}{(2\pi\sigma_j^2)^{\frac{d}{2}}} e^{-\frac{1}{2\sigma_j^2}\|x_i-\mu_j\|_2^2}}{\sum\limits_{\ell=0}^{k} \frac{\eta_\ell}{(2\pi\sigma_\ell^2)^{\frac{d}{2}}} e^{-\frac{1}{2\sigma_\ell^2}\|x_i-\mu_\ell\|_2^2}}$$

  $$\lim_{\sigma_j^2 \to 0} \gamma_{ij} = \begin{cases} 1 & \text{if } j = \underset{j \in \{1,\dots,k\}}{\arg\min} \|x_i - \mu_j\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

  M-step:

  $$\mu_j^{(t+1)} \leftarrow \frac{\sum\limits_{i=1}^{n} \gamma_{ij} X_i}{\sum\limits_{i=1}^{n} \gamma_{ij}} = \frac{\sum\limits_{i=1}^{n} \mathbb{1}(Z_i = j) X_i}{\sum\limits_{i=1}^{n} \mathbb{1}(Z_i = j)}$$