Sprig '21 **Mixture Models** CS181

Data $D = \{\underline{x}_1, \ldots, \underline{x}_n\}$ No target labels

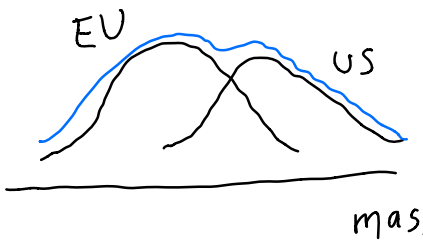
Last lecture

1) K-means clustering
2) HAC

arguably a bit ad hoc, HAC is flexible but less useful in high dimensions

Mixture models

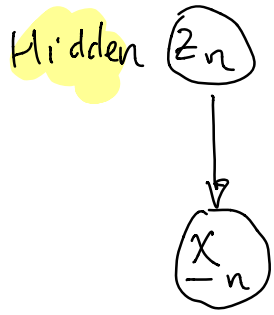
Probabilistic, generative view. Data $\{\underline{x}_n\}$ as coming from a mixture of components of a distribution.



Use D to estimate parameters of a mixture model

cluster ≡ component

for $\underline{x}$, cluster according to most likely component



cluster label — $p(z)$ — eg., location

data — $p(\underline{x}|z)$ — eg., mass of a car in a location

$$p(x,z) = p(z)\,p(\underline{x}|z)$$

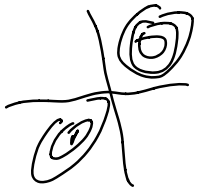
Predict $p(z|x) \propto p(z)\,p(\underline{x}|z)$

Gaussian Mixture Model (GMM)

Classes $C_1, \ldots, C_K$ $\quad \underline{x}_n \in \mathbb{R}^D$

$p(z = C_k) = \theta_k$ categorical

$p(\underline{x}|z = C_k) = N(\underline{x}; \underline{\mu}_k, \Sigma_k)$



Parameters $\underline{w}$: $\{\theta_k\}_{k=1}^K$ $\{\underline{\mu}_k, \underline{\Sigma}_k\}_{k=1}^K$

Max. Lik. Est.

$\underline{\mu}_k$ ↑ mean — $\underline{\Sigma}_k$ ↑ covar

log likelihood $\mathcal{L}(X; \underline{w}) = \sum_n \ln p(\underline{x}_n)$

$$= \sum_n \ln \sum_k p(\underline{x}_n, C_k)$$

$$= \sum_n \ln \sum_k p(C_k)\,p(\underline{x}_n|C_k)$$

$$= \sum_n \ln \sum_k \theta_k \, N(\underline{x}_n ; \underline{\mu}_k, \underline{\Sigma}_k)$$

No analytical solution!
Log-sum structure prevents decomposing over the parameters.

Consider a simpler, "complete data" problem. Suppose we observe class labels $z_1, \ldots, z_N$.

Complete-data log likelihood (z_n 1-hot)

$$\ell(X, Z; \underline{w}) = \sum_n \ln p(\underline{x}_n, z_n)$$
$$= \sum_n \ln p(z_n)\, p(\underline{x}_n | z_n)$$
$$= \sum_n \ln p(z_n) + \sum_n \ln p(\underline{x}_n | z_n)$$

$$p(z_n) = \prod_k \theta_k^{z_{nk}} \qquad p(\underline{x}_n | z_n) = \prod_k N(\underline{x}_n ; \underline{\mu}_k, \underline{\Sigma}_k)^{z_{nk}}$$

$$\ell(X, Z; \underline{w}) = \sum_n \sum_k z_{nk} \ln \theta_k + \sum_n \sum_k z_{nk} \ln N(\underline{x}_n ; \underline{\mu}_k, \underline{\Sigma}_k)$$

Solve analytically

$$\hat{\theta}_k = \frac{N_k}{N} \qquad \hat{\underline{\mu}}_k = \text{mean}_{n \text{ s.t. } z_n = C_k}(\underline{x}_n) \qquad \hat{\underline{\Sigma}}_k = \text{var}_{n \text{ s.t. } z_n = C_k}(\underline{x}_n)$$

N_k = # examples in C_k

Expectation-Maximization Algorithm

Iterative approach!

Initialize parameters $\{\theta_k\}$ $\{\mu_k, \Sigma_k\}$, repeat

1. E-step. Use parameters to predict distribution on assignments z_n for each example
2. M-step. Update parameters, maximizing expected complete data log likelihood given predicted assignments.

E-M algorithm. Very powerful, general method to maximize likelihood for models with latent variables (eg., assignments z).

M-step often has an analytical solution.

E-step Soft class assignment

$$p(z_n = C_k \mid \underline{x}_n) = q_{nk} \propto \theta_k \mathcal{N}(\underline{x}_n ; \underline{\mu}_k, \underline{\Sigma}_k)$$

M-step Work with the expected complete-data log likelihood:

$$\max_{\underline{w}} \mathbb{E}_{Z \sim q}\left[L(X, Z ; \underline{w}) \right]$$

Analytical solution:

$$\hat{\theta}_k \leftarrow \frac{N_k}{N} \qquad \hat{\underline{\mu}}_k = \frac{1}{N_k} \sum_n q_{nk} \underline{x}_n$$

$$N_k = \sum_n q_{nk} \qquad \hat{\underline{\Sigma}}_k = \frac{1}{N_k} \sum_n q_{nk} (\underline{x}_n - \underline{\mu}_k)(\underline{x}_n - \underline{\mu}_k)^T$$

Suppose
$\theta_1 = \theta_2 = \frac{1}{2}$
$\underline{\mu}_1 = \underline{\mu}_2$
$\underline{\Sigma}_1 = \underline{\Sigma}_2$

Properties

At each step, E-M algorithm adjusts parameters to improve likelihood of observed data D

$$p(D ; \underline{w}^{(0)}) < p(D ; \underline{w}^{(1)}) < \cdots$$

$\underline{w}^{(t)}$: parameters in step t

→ Converges to a local optimum (Restarts can help)

Understanding M-step

$$p(\underline{x}_n) = \sum_k p(\underline{x}_n, C_k) = \sum_k p(z_n = C_k)\, p(\underline{x}_n \mid z_n = C_k)$$

Observed log likelihood

$$\ln p(\underline{x}_n) = \ln \sum_k \theta_k \mathcal{N}(\underline{x}_n ; \underline{\mu}_k, \underline{\Sigma}_k)$$

Complete-data log lik:

$$\begin{aligned}
\ln p(\underline{x}_n, z_n) &= \ln p(z_n)\, p(\underline{x}_n \mid z_n) \\
&= \ln p(z_n) + \ln p(\underline{x}_n \mid z_n) \\
&= \ln \prod_k \theta_k^{z_{nk}} + \ln \prod_k \mathcal{N}(\underline{x}_n ; \underline{\mu}_k, \underline{\Sigma}_k)^{z_{nk}} \\
&= \sum_k z_{nk} \ln \theta_k + \sum_k z_{nk} \ln \mathcal{N}(\underline{x}_n ; \underline{\mu}_k, \underline{\Sigma}_k)
\end{aligned}$$

Expected complete data log lik:

$$\begin{aligned}
\mathbb{E}_{z_n} \ln p(\underline{x}_n, z_n) &= \sum_k q_{nk} \ln p(C_k)\, p(\underline{x}_n \mid C_k) \\
&= \sum_k q_{nk} \ln \theta_k + \sum_k q_{nk} \ln \mathcal{N}(\underline{x}_k ; \underline{\mu}_k, \underline{\Sigma}_k)
\end{aligned}$$

M-step works with

$$\mathbb{E}_z \ln \prod_n [\cdot] = \mathbb{E}_z \sum_n \ln [\cdot] = \sum_n \mathbb{E}_{z_n} \ln [\cdot]$$

Notes

① Initialization matters!

eg, what happens if

(a) $\hat{\theta}_k = \frac{1}{K}$ all k, all $\hat{\underline{\Sigma}}_k$, $\hat{\underline{\mu}}_k$ same each k

(b) $\hat{\theta}_1 > \{\hat{\theta}_2, \ldots, \hat{\theta}_K\}$,

all $\hat{\underline{\Sigma}}_k$, $\hat{\underline{\mu}}_a$ same each k

② Choose a variation, eg.

	θ	$\underline{\mu}$	$\underline{\Sigma}$
general $N(\underline{x}; \underline{\mu}_k, \underline{\Sigma}_k)$	$K-1$	KD	KD^2
diagnol $N(\underline{x}; \underline{\mu}_k, \underline{\Sigma}_k)$ (↑ diag.)	$K-1$	KD	KD
spherical (isotropic) $N(\underline{x}; \underline{\mu}_k, \sigma^2 \underline{I})$	$K-1$	KD	K

(linear decision boundaries)

Param.

(3) Connects to K-means!

K-means Repeats

1. Hard assign each example to nearest prototype μ_k
2. Update prototype to centroid of assigned points (mean)

E-M Fix class prob at $1/K$ with

Consider spherical Gaussian, ~~some~~ constant εI for $\varepsilon > 0$, small.

E-step: Hard assign

M-step: Update μ_k to new centroids of assigned data.