Alpaydin Ch 7

Clustering

Mixtures - E.G. 7 and 7

$$p(\underline{x}) = \sum_{t=1}^{k} p(\underline{x} \mid G_t) P(G_t)$$

G_t are the mixture components

$p(\underline{x} \mid G_t)$ component densities. Different Classes

$P(G_t)$ mixture proportion.

If component densities are Gaussian we have $p(\underline{x} \mid G_t) \sim N(\underline{\mu}_i, \underline{\Sigma}_i)$

$\Phi = \{ P(G_t), \underline{\mu}_t, \underline{\Sigma}_t \}_{t=1}^{k}$ are parameters that need to be estimated.

In supervised case, we know the classes, and the assignment of instances to classes.

$$r_i^t = \begin{cases} 1, & \text{if } x^t \in C_i \\ 0, & \text{otherwise} \end{cases}$$

In this case, it is straightforward to learn the parameters

$$\hat{P}(G_t) = \sum_t r_i^t / N$$

$$\underline{m}_t = \sum_t r_i^t \underline{x}^t / \sum_t r_i^t$$

$$\underline{S}_t = \sum_t r_i^t (\underline{x}^t - \underline{m}_i)(\underline{x}^t - \underline{m}_i)^T / \sum_t r_i^t$$

Mixture Densities (cont.)

But, usually we do not know the classes.
This is called unsupervised learning
ie. r^t is not known.

We need to estimate r^t — later, we show that the EM algorithm can be used.

k-means clustering.

Vector quantization:

Sample $\mathcal{X} = \{\underline{x}^t\}_{t=1}^N$

k reference vectors $\underline{m}_j \quad j = 1 \text{ to } k.$

Represent $\underline{x}^t$ by the most similar entry $\underline{m}_i$

$$\|\underline{x}^t - \underline{m}_i\| = \min_j \|\underline{x}^t - \underline{m}_j\|$$

$\underline{m}_i$ — codebook vectors.

Compression — results from quantization.

How to calculate the best $\{\underline{m}_j\}$?

(3)

K-means (cont'd)

Reconstruction Error

$$E(\{\underline{m}_i\}_{i=1}^k \mid X) = \sum_t \sum_i b_i^t \|\underline{x}^t - \underline{m}_i\|^2$$

where $b_i^t = \begin{cases} 1, & \text{if } \|\underline{x}^t - \underline{m}_i\| = \min_j \|\underline{x}^t - \underline{m}_j\| \\ 0, & \text{otherwise} \end{cases}$

k-means is an iterative procedure that seeks to minimize $E(\cdot)$.

Start with random $\underline{m}_i$.

Then compute b_i^t

re-estimate the means $\underline{m}_i = \frac{\sum_t b_i^t \underline{x}^t}{\sum_t b_i^t}$.

Repeat.

This procedure is guaranteed to converge (almost always).

But the final results depend on the initializations.

Alternative Initialization Strategies:

- Take random k instances as initial $\underline{m}_i$
- Calculate the means of all the data, then add small perturbations
- Calculate the principal components, divide its range into k intervals, etc.

(4)

Expectation-Maximization (EM) Algorithm.

$$\mathcal{L}(\Phi|X) = \log \prod_t p(x^t|\Phi)$$
$$= \sum_t \log \sum_{i=1}^k p(x^t|G_i) P(G_i)$$

EM Introduce hidden variable Z which associate the data to the class.

$$Z^t = (z_1^t, \ldots, z_k^t)$$ $z_i^t = 1$ if x^t belong to class G_i, 0, otherwise.

$$p(z^t) = \prod_{i=1}^k \pi_i^{z_i^t}$$

$$p(x^t|z^t) = \prod_{i=1}^k p_i(x^t)^{z_i^t}$$ $p_i(x^t)$ short for $p(x^t|G_i)$

$$p(x^t, z^t) = p(z^t) p(x^t|z^t)$$

$$\mathcal{L}_c(\Phi|X,Z) = \log \prod_t p(x^t, z^t|\Phi)$$
$$= \sum_t \log p(x^t, z^t|\Phi)$$
$$= \sum_t \{\log p(z^t|\Phi) + \log p(x^t|z^t, \Phi)\}$$
$$= \sum_t \sum_i z_i^t \{\log \pi_i + \log p_i(x^t|\Phi)\}$$

(5) EM cont'd

E-step

$$Q(\phi \mid \phi^{(l)}) = E[\log p(X, Z) \mid X, \phi^{(l)}]$$
$$= E[L_c(\phi \mid X, Z) \mid X, \phi^{(l)}]$$
$$= \sum_t \sum_i E[z_i^t \mid X, \phi^{(l)}] \{\log \pi_i + \log p_i(x^t \mid \phi_i)\}$$

where $E[z_i^t \mid X, \phi^{(l)}]$

$$= E[z_i^t \mid x^t, \phi^{(l)}]$$
$$= P(z_i^t = 1 \mid x^t, \phi^{(l)})$$
$$= \frac{p(x^t \mid z_i^t = 1, \phi^{(l)}) P(z_i^t = 1 \mid \phi^{(l)})}{p(x^t \mid \phi^{(l)})}$$ Bayes rule

$$= \frac{p_i(x^t \mid \phi^{(l)}) \pi_i}{\sum_j p_j(x^t \mid \phi^{(l)}) \pi_j}$$

$$= \frac{p(x^t \mid G_i, \phi^{(l)}) P(G_i)}{\sum_j p(x^t \mid G_j, \phi^{(l)}) P(G_j)}$$

$$= P(G_i \mid x^t, \phi^{(l)}) = h_i^t$$

Expected value of hidden variable, $E[z_i^t]$ is the posterior probability that x^t is generated by G_i.

"soft label"

(6) EM (Cont)

M-step: maximize to get next of parameter $\phi^{l+1} = \arg\max_{\phi} Q(\phi|\phi^l)$

$$Q(\phi|\phi^j) = \sum_t \sum_i h_i^t \{\log \pi_i + \log p_i(x^t|\phi^l)\}$$

$$= \sum_t \sum_i h_i^t \log \pi_i + \sum_t \sum_i h_i^t \log p_i(x^t|\phi^l)$$

solve to get $\pi_i = \sum_t h_i^t / N$

If we assume Gaussian components $p_j(\underline{x}^t|\phi) \sim N(\underline{m}_i, \underline{S}_i)$

M-step is $\underline{m}_i^{l+1} = \sum_t h_i^t \underline{x}^t / \sum_t h_i^t$

$$\underline{S}_i^{l+1} = \sum_t h_i^t (\underline{x}^t - \underline{m}_i^{l+1})(\underline{x}^t - \underline{m}_i^{l+1})^T / \sum_t h_i^t$$

where $h_i^t = \dfrac{\pi_i |\underline{S}_i|^{-1/2} \exp\{-\frac{1}{2}(\underline{x}^t - \underline{m}_i)^T \underline{S}_i^{-1}(\underline{x}^t - \underline{m}_i)\}}{\sum_j \pi_j |\underline{S}_j|^{-1/2} \exp\{-\frac{1}{2}(\underline{x}^t - \underline{m}_j)^T \underline{S}_j^{-1}(\underline{x}^t - \underline{m}_j)\}}$

(7)

Supervised Learning after Clustering.

Clustering can be used for data exploration by grouping instances.

If such groups are found, they can be named and their attributes defined.

Clustering is frequently used as a pre-processing stage.

(8) # Modern View of EM.

Task estimate parameters ϕ from $p(x, z \mid \phi)$.

Problem we can't observe z

Proposed Solution: Estimate ϕ from $p(x \mid \phi)$ where $p(x \mid \phi) = \sum_z p(x, z \mid \phi)$

Use iterative algorithm, EM, to estimate ϕ.

Formulate minimize $-\log p(x \mid \phi)$ w.r.t. ϕ.

Equivalent to minimize

$$F[\phi, q(z)] = -\log p(x \mid \phi) + \sum_z q(z) \log \frac{q(z)}{p(z \mid x, \phi)}$$

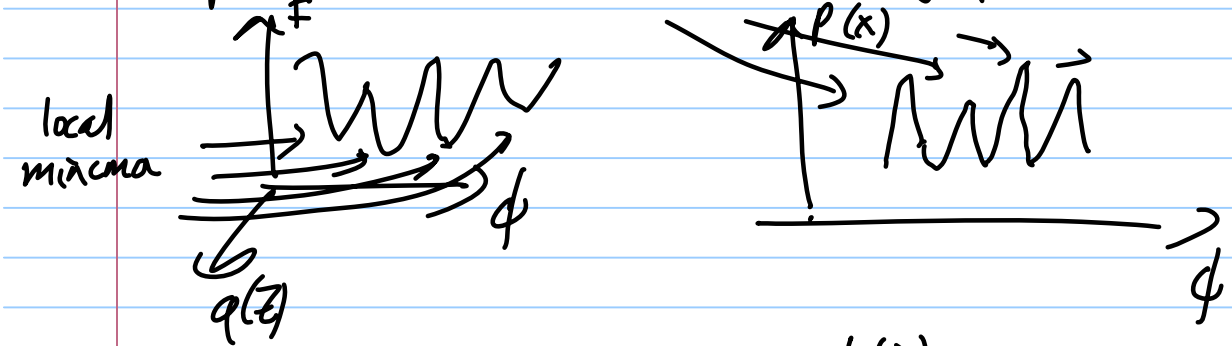
w.r.t. ϕ & $q(z)$.

Why → because the second term is ≥ 0 and $= 0$ only if $q(z) = p(z \mid x, \phi)$

Kullback-Leibler divergence.

(9) Algorithm (EM) minimize $F[\phi, q(z)]$ wrt ϕ & $q(z)$ alternately.

This is guaranteed to always decrease F and converge to a local minimum. This corresponds to a local maximum of $p(x|\phi)$



Initialize the parameters $\phi^{(0)}$, then repeat the two steps: at time t.

E-step: minimize F w.r.t. $q(z)$ gives

$$q^{t+1}(z) = p(z|x, \phi^t)$$

M-step: minimize F w.r.t. ϕ gives

$$\phi^{t+1}(z) = \arg\min_{\phi} -\sum_z q^{t+1}(z) \log p(x, z|\phi)$$

see next page ...

(10) E-step follows from standard expression of $F[\phi, q(z)] = -\log p(x|\phi) + \underbrace{\sum_z q(z) \log \left\{ \frac{q(z)}{p(z|x,\phi)} \right\}}$

minimize by setting $q(z) = p(z|x,\phi)$.

M-step follows by re-expressing

$$F = -\log p(x|\phi) + \sum_z q(z) \log q(z) - \sum_z q(z) \log p(z|x,\phi)$$

$$F = \sum_z q(z) \log(z) - \underbrace{\sum_z q(z) \log p(x|\phi) - \sum_z q(z) \log p(z|x,\phi)}$$

using $\sum_z q(z) = 1$

$$-\sum_z q(z) \log \{ p(x|\phi) p(z|x,\phi) \}$$

$$-\sum_z q(z) \log p(x,z|\phi) \quad //$$

This is the "modern" formulation of EM. It gives a proof of convergence. It shows where the two steps come from.

It also leads to variants of EM (beyond scope of the course).