COMP3206: Classification: generative approach

Srinandan Dasmahapatra

2017/8

Generation by sequence of noisy machines

- A **generative model** is one that provides a (stochastic) mechanism for reproducing the statistical properties of the observed data.
- A binomial distribution summarises the statistics of multiple binary outcomes, (Bernoulli trials). Example: biased coin tosses. A multinomial distribution captures the statistics of a loaded die.
- Model acceptance of applicant to ECS as Bernoulli trial – either you get accepted or not. Table of **acceptance rates**:

	ECS	E	CS
Male	28%	$\theta^{E}_{M} = 40\%$	$\theta^{CS}_{M} = 10\%$
Female	23%	$\theta^{E}_{F} = 50\%$	$\theta^{CS}_{F} = 20\%$

- **Application rates** $(\pi^{E}_{F}, \pi^{E}_{M}, \pi^{CS}_{F}, \pi^{CS}_{M}) = (\overline{.1, .6}^{E}, \overline{.9, .4}^{CS})$
- Overall acceptance rate for $\alpha = M, F$ is $\pi^{E}_{\alpha}\theta^{E}_{\alpha} + \pi^{CS}_{\alpha}\theta^{CS}_{\alpha}$.

Mixture distributions are generative models

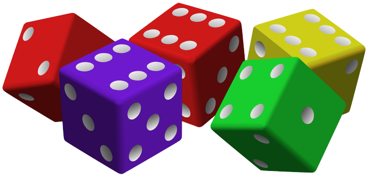
- A **generative model** is one that provides a (stochastic) mechanism for reproducing the statistical properties of the observed data.
- The stochastic mechanism might involve a sequence of random choices, each modelled by a (simpler) stochastic mechanism
- Previous example: A model for choice followed by model for acceptance.
- A model of acceptance (M::F) is refined by introducing a new variable (M(E)::F(E), M(C)::F(C)).
- Mixture distributions refine models by introducing variables that can be unobservable.

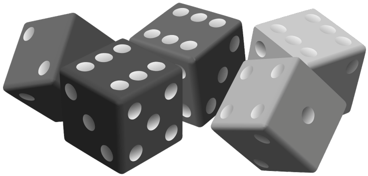
Mixture of multinomials: rolling several loaded dice

- The probability of $x = (x^{(1)}, \ldots, x^{(n)}, \ldots, x^{(N)})$ outcomes of N rolls of a die is

$$p_d(x|\boldsymbol{\theta}) = \left(\frac{N!}{n_1!n_2!n_3!n_4!n_5!n_6!}\right) \theta_1^{n_1}\theta_2^{n_2}\theta_3^{n_3}\theta_4^{n_4}\theta_5^{n_5}\theta_6^{n_6}.$$

  n_j counts occurrences of j.
- We have seen how to estimate $\boldsymbol{\theta}$ from observations by MLE (counts) or Bayesian posteriors (counts + pseudocounts).
- But what if each throw is from one of several dice that look the same?





Mixture of multinomials: rolling several loaded dice

- If x describes rolled outcomes of H identical looking but differently weighted dice, chosen with probability π_i, called a mixture distribution, with mixture weights $\pi_i \geqslant 0$ and $\sum_i \pi_i = 1$.

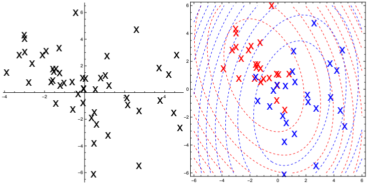
$$p(x) = \sum_{i=1}^{H} p(x|\boldsymbol{\theta}^{(i)})\pi_i$$

- **Estimation**: infer probabilities $\boldsymbol{\theta}^{(i)}$ of dice i and mixture weights π_i from data x.
- **Classification**: which die is the most probable generator of observation $X = x^k \in \{1, \cdots, 6\}$?

$$\operatorname*{argmax}_i P(i|X = x^k) = \frac{P(X = x^k|\boldsymbol{\theta}^{(i)})\pi_i}{P(X = x^k)}$$

- Infer sequences $(i_1, i_2, i_3, \ldots)$ by hidden Markov models (HMMs)

Mixture of Gaussians: un/supervised learning



- **Estimation**: (input) data $\mathcal{X} = \{x^1, \ldots, x^N\}$ (left); (output) colours and mathematical description of contours (right). That is, find $\boldsymbol{\theta}^{(i)}$: component means $\boldsymbol{\mu}_i$, covariances $\boldsymbol{\Sigma}_i$ and mixture weights π_i.
- **Classification**: for which assignment of colour is the probability of observations maximised?

$$\operatorname*{argmax}_i P(i|\mathbf{X} = x^k) = \frac{P(\mathbf{X} = x^k|\boldsymbol{\theta}^{(i)})\pi_i}{P(\mathbf{X} = x^k)}, \; \boldsymbol{\theta}^{(i)} = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i \in \{\text{red},$$

- **Clustering**: no coloured training examples (**unsupervised learning**).

Mixture of Gaussians

- For random variable $\mathbf{X}$ described by Gaussian pdf $\mathcal{N}(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, p-dimensional data point $x^n \in (x, x + dx)$ has probability density $p(\mathbf{X} = x)dx$ where

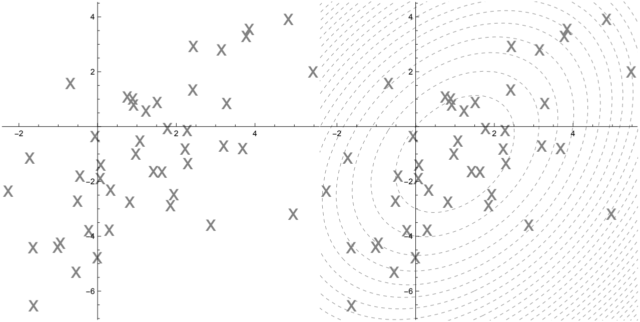
$$p(\mathbf{x}^n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\sqrt{(2\pi)^p|\boldsymbol{\Sigma}|})^{-1} \exp\left(-\frac{1}{2}(\mathbf{x}^n - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} -^n \boldsymbol{\mu})\right),$$

  where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. A mixture of Gaussians has a probability density

$$p(\mathbf{X} = \mathbf{x}^n) = \sum_{i=1}^{H} \mathcal{N}(\mathbf{X} = \mathbf{x}^n|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\pi_i$$

  where the mixture weights $\pi_i \geqslant 0$ satisfy $\sum_i \pi_i = 1$,.
- Need to learn the component means $\boldsymbol{\mu}_i$, covariances $\boldsymbol{\Sigma}_i$ and mixture weights π_i from data $\mathcal{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^n, \ldots, \mathbf{x}^N\}$.

Learning the parameters of a Gaussian by Maximum Likelihood



Estimation: (input) data $\mathcal{X} = \{x^1, \ldots, x^N\}$ (left); (output) mathematical description of contours (right). That is, find mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

Maximum Likelihood Estimation (MLE) of parameters of Gaussian distribution

- Assuming data are drawn i.i.d. from Gaussian $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda})$, $\boldsymbol{\Lambda} \triangleq \boldsymbol{\Sigma}^{-1}$, the log likelihood $\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ is

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = -\sum_{n=1}^{N} \frac{1}{2}(\mathbf{x}^n - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\mathbf{x}^n - \boldsymbol{\mu}) + \frac{N}{2} \log\det(\boldsymbol{\Lambda}) + \text{const.}$$

- Optimal $\boldsymbol{\mu}$: $\frac{\partial}{\partial \boldsymbol{\mu}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = 0 = \sum_n \boldsymbol{\Lambda} (\mathbf{x}^n - \boldsymbol{\mu})$

$$\sum_n \boldsymbol{\Lambda} \mathbf{x}^n = \sum_n \boldsymbol{\Lambda} \boldsymbol{\mu} = N \boldsymbol{\Lambda} \boldsymbol{\mu} \Rightarrow \boldsymbol{\mu} = \frac{1}{N} \sum_n \mathbf{x}^n.$$

- Optimal $\boldsymbol{\Lambda}$: $\frac{\partial}{\partial \boldsymbol{\Lambda}} \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = 0$. ($\log\det A = \operatorname{trace} \log A$ for any $\mathbf{A}$.)

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}^n - \boldsymbol{\mu})(\mathbf{x}^n - \boldsymbol{\mu})^T.$$

Learning the parameters of a Mixture of Gaussians by Maximum Likelihood

- Cannot maximise log likelihood in the same way:

$$\log p(\mathbf{x}) = \log \left(\sum_{i=1}^{H} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\pi_i \right)$$

- Introduce a hidden variable z_i that takes a value $0/1$ and $\sum_{i=1}^{H} z_i = 1$. The marginal distribution over $\mathsf{Z} = (Z_1, \ldots, Z_i, \ldots, Z_H)$ is specified as the mixing weights π_i, *i.e.*, $p(Z_i = z_i = 1) = \pi_i$.
- Define joint distribution over the hidden and observed variables (X, Z).

Introducing "responsibilities" of each mixture component by Bayes

- The posterior distributions over the hidden variables $\gamma(Z_i) \equiv p(Z_i = 1|\mathbf{X})$ are called *responsibilities* and are obtained from the joint using Bayes' rule:

$$\gamma(Z_i) = \frac{p(\mathbf{x}|Z_i = 1)p(Z_i = 1)}{\sum_j p(\mathbf{x}|Z_j = 1)p(Z_j = 1)} = \frac{\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\pi_i}{\sum_j \mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\pi_j}$$

Learning the parameters of a Mixture of Gaussians by Expectation Maximisation: $\boldsymbol{\mu}_k$ from $\frac{\partial}{\partial \boldsymbol{\mu}_k}\mathcal{L} = 0$

- At a maximum of log likelihood:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_n \log p(\mathbf{x}^n) = \sum_n \frac{\partial}{\partial \boldsymbol{\mu}_k} \log \left(\sum_{i=1}^{H} \mathcal{N}(\mathbf{x}^n | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \pi_i \right) \\
&= -\sum_{n}^{N} \overbrace{\frac{\mathcal{N}(\mathbf{x}^n | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \pi_i}{\sum_{j=1}^{H} \mathcal{N}(\mathbf{x}^n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \pi_j}}^{\text{responsibility } \gamma(Z_{ni})} \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}^n - \boldsymbol{\mu}_i) \\
&= \boldsymbol{\Sigma}_i^{-1} \left(\sum_{n}^{N} \gamma(Z_{ni}) \mathbf{x}^n - \boldsymbol{\mu}_i \sum_n \gamma(Z_{ni}) \right),
\end{aligned}
$$

where $\gamma(Z_{nj})$ is the responsibility of the j-th mixture component for the n-th data point.

Learning the parameters: the means from $\frac{\partial}{\partial \boldsymbol{\mu}_k}\mathcal{L} = 0$

- From the optimisation in the previous slide:

$$\boldsymbol{\mu}_k = \frac{\sum_n \gamma(Z_{nk})\mathbf{x}^n}{\sum_n \gamma(Z_{nk})} = \frac{1}{N_k}\sum_n \gamma(Z_{nk})\mathbf{x}^n$$

Defined N_k as the accumulated contribution to all data points of each mixture component (accumulated responsibilities).

Learning the parameters of a Mixture of Gaussians: $\boldsymbol{\Sigma}_k$, π_k

- Similarly, by taking derivatives w.r.t. $\boldsymbol{\Sigma}_k^{-1}$,

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_n \gamma(Z_{nk})(\mathbf{x}^n - \boldsymbol{\mu}_k)(\mathbf{x}^n - \boldsymbol{\mu}_k)^T$$

- For π_k: Constrained optimisation with cost function $\log p(\mathbf{x}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^{H} \pi_k - 1 \right)$:

$$\pi_k = \frac{N_k}{\sum_k N_k} = \frac{N_k}{N}$$

- We started with some $\{\pi, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ and computed responsibilities $\gamma(Z_{ni})$. These in turn determine a new set of values for $\{\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$.

Learning the parameters of a Mixture of Gaussians by Expectation Maximisation: $\boldsymbol{\Sigma}_k$, π_k

- We started with some $\{\pi, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ and computed responsibilities $\gamma(Z_{ni})$. These in turn determine a new set of values for $\{\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$.
- **Expectation Maximisation** iteratively estimates parameters of each mixture component by taking:
  - expectation values (E step) of hidden and visible data with respect to current distribution, called total data likelihood.
  - perform maximum likelihood (M step): counting step – counts occurrence of fractions (estimated as responsibilities) of the event type.

Generative mixture model: Probabilistic Latent Semantic Analysis (PLSA) for text classification - I

- Document corpus $D \ni d$ contains words $w \in W$. Assume $t \in T$ hidden topics to explain co-occurrence (w, d), $0 \leqslant \pi_t \leqslant 1$, with $\sum_{t=1}^{|T|} \pi_t = 1$.
- For n^{th} word position in document d draw topic $t_n \sim p(Z_n = t_n \in T | d \in D)$
- Given topic for this word position, draw word $w_n \sim p(W_n = w_n | Z_n = t_n \in T) = \theta_{w_n|t}$
- If document d has $|d|$ words,

$$p(w_1, \ldots, w_{|d|}, Z_1, \ldots, Z_{|d|}) = \prod_{n=1}^{|d|} p(w_n|Z_n)p(Z_n|d)$$

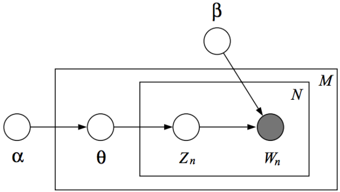
Generative mixture model: Probabilistic Latent Semantic Analysis (PLSA) for text classification - 2

- Vocabulary W; $t \in T$ hidden topics; $d \in D$ document corpus
- How to best explain co-occurrence (w, d), $0 \leqslant \pi_t \leqslant 1$, with $\sum_{t=1}^{|T|} \pi_t = 1$?
- To generate word in nth location in document d, draw topic $t \sim p(Z_n = t \in T | d \in D)$, and draw word $w_n \sim p(W_n = w_n | Z_n = t \in T) = \theta_{w_n|t}$. Each document has one topic.
- Marginal distribution over observed data:

$$
\begin{aligned}
p(\{w_n\}) &= \sum_{\{Z_n = t_n\}} \prod_{n=1}^{|d|} p(w_n | Z_n = t_n) p(Z_n = t_n | d) \\
&= \prod_{n=1}^{|d|} \sum_{t=1}^{|T|} p(w_n | t_n) p(t_n | d)
\end{aligned}
$$

Topic modelling with Latent Dirichlet Allocation

- Corpus of documents, size M. Specific document, size N.
- For each word position n, word w_n has topic $Z_n = t_n$ generating it. Multinomial: $w_n \sim p(W_n = w_n \in W | Z_n = t_n, \theta)$.



- For each word a new topic is chosen, not one topic for the entire document. $Z_n \sim \theta$. Multinomial: $p(Z_n = t_n \in T | \theta)$. Mixed membership model.
- Prob(topic) $\theta \sim$ Dirichlet(α) (prior). Prob(word) has β prior.

Generative classifiers

- Introduced estimation of parameters of probability distributions in classification context
- Bayes' theorem provides method of classification
- Estimation performed by MLE (or Bayesian posterior distributions)
- Mixture models estimated by iterative procedure, EM.
- Examples of discrete mixture distributions in topic modelling.
- Next subject: discriminative classifiers