Advanced Machine Learning Subsidary Notes

Lecture 18: Generative Models

Adam Prügel-Bennett

May 2, 2020

1 Keywords

- Generative models, graphical models, LDA

2 Main Points

2.1 Bayesian Inference

- Most Bayesian inference involves constructing a model of the underlying data generation process and using Bayes' rule to learn unknown properties of the model
- In building models we use random variables, X, Y, Z, etc. to model quantities we are uncertain about
- We associate *probability mass functions* $\mathbb{P}[X, Y, Z]$ (for discrete random variable) or *probability densities* $f_{X,Y,Z}(x, y, z)$ (for continuous random variables)
- Often our tasks will be to infer these probability distributions (or parameters of these probability distribution) for quantities of interest
- In classical machine learning we may think the feature vector $\boldsymbol{X}$ as being a random variable and the prediction Y as being a second random variable
- **Discriminative Models**
  - Often our goal is to learn the probability distribution $\mathbb{P}[Y|\boldsymbol{X}]$
  - Very often we would parameterise this distribution with some parameters $\boldsymbol{\Theta}$ and our task would be to learn these parameters based on training data
- **Generative Models**
  - Surprisingly it is often easier to model the joint probability $\mathbb{P}[Y, \boldsymbol{X}]$
  - This means that we model the process of both generating the targets and the feature vectors together
  - These are known as *generative models* as they allow us to generate random examples
  - We don't necessary want to use them to generate random samples it just makes the modelling process easier (although you need to get used to this as it feel counter-intuitive)
  - we can use generative models to do discrimination
  - Examples of generative models include *Hidden Markov Models* and *Topic Models* (covered later)
- **Latent Variables**
  - In building probabilistic models we often model quite complicated processes

- To do this we often introduce intermediate processes
- This leads to introduce other random variables that we actually never observe
- These are known as **latent variable**
- Often our model will involve many different layers between the inputs $\boldsymbol{X}$ and targets Y: this process is sometimes known as *hierarchical modelling*

- **Mixtures of Gaussians**
  - To illustrate latent variables and a simple hierarchical model we consider a classic probabilistic model known as *mixture of Gaussians*
  - We consider a concrete scenario
  - We suppose we are observing the decay of two types (A and B) of short-lived particles
  - We can measure their half lives, X_i, but we don't know the type of particle
  - We have a measurement error of the half-life
  - Let $Z_i \in \{0,1\}$ equal 1 if particle i is of type A and 0 if it is of type B
  - The probability distribution of the half-life measurement is therefore

$$f(X_i|Z_i,\boldsymbol{\Theta}) = Z_i\,\mathcal{N}\big(X_i\big|\mu_A,\sigma_A^2\big) + (1-Z_i)\,\mathcal{N}\big(X_i\big|\mu_B,\sigma_B^2\big)$$

    * where μ_A and μ_B are the expected half-lives for particles of type A and B respectively
    * σ_A and σ_B are the standard deviations in the measurements
    * this just says that if the i^{th} particle is of type A then the probability of X_i is $\mathcal{N}\big(X_i\big|\mu_A,\sigma_A^2\big)$ and if it is of type B it is $\mathcal{N}\big(X_i\big|\mu_B,\sigma_B^2\big)$
  - We assume that we have m observations (e.g. $m = 1\,000$)

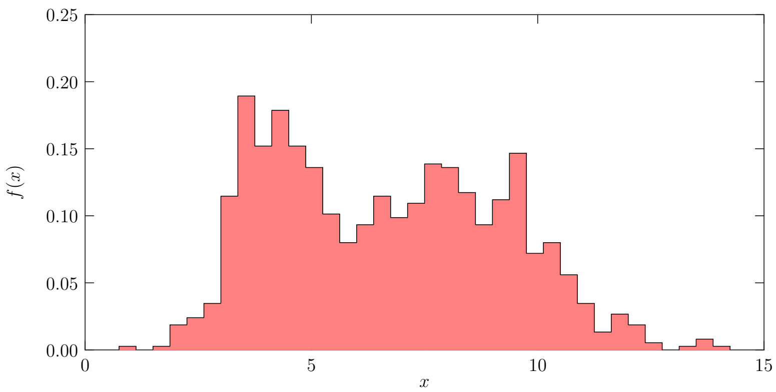

Figure 1: Example of distribution of half-lives

  - Our job is to infer the random variables $\boldsymbol{\Theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2, p)$, where $p = \mathbb{P}[Z_i = 1]$ is the probability of the particle being type A
  - We can do a full Bayesian calculation, but let us just use a maximum likelihood
  - The maximum likelihood of the data $\mathcal{D} = \{X_i | i = 1, 2, \ldots, m\}$ is

$$\begin{aligned} f(\mathcal{D}|\boldsymbol{\Theta}) &\overset{(1)}{=} \sum_{\boldsymbol{Z}\in\{0,1\}^m} f(\mathcal{D},\boldsymbol{Z}|\boldsymbol{\Theta}) \\ &\overset{(2)}{=} \prod_{i=1}^{m} \sum_{Z_i\in\{0,1\}} f(X_i, Z_i|\boldsymbol{\Theta}) \overset{(3)}{=} \prod_{i=1}^{m} \sum_{Z_i\in\{0,1\}} f(X_i|Z_i,\boldsymbol{\Theta})\,\mathbb{P}[Z_i] \end{aligned}$$

1. where we marginalise out the latent variables $\boldsymbol{Z} = (Z_1, Z_2, \ldots Z_n)$
2. we assume the data is independent

3. we use the identity $f(X_i, Z_i|\boldsymbol{\Theta}) = f(X_i|Z_i, \boldsymbol{\Theta})\, \mathbb{P}[Z_i]$

- It is usually easier working with the log-likelihood

$$\begin{aligned}\log\bigl(f(\mathcal{D}|\boldsymbol{\Theta})\bigr) &= \sum_{i=1}^{m} \log\bigl(f(X_i|Z_i = 1)\, \mathbb{P}[Z_i = 1] + f(X_i|Z_i = 0)\, \mathbb{P}[Z_i = 0]\bigr) \\ &= \sum_{i=1}^{m} \log\bigl(p\, \mathcal{N}\bigl(X_i \big| \mu_A, \sigma_A\bigr) + (1-p)\, \mathcal{N}\bigl(X_i \big| \mu_B, \sigma_B\bigr)\bigr)\end{aligned}$$

- We could do gradient descent on this, but it is an ugly expression to work with

- **Expectation Maximisation**
  - Rather than maximise the likelihood directly we iteratively maximise the expected log-likelihood starting form some guess $\boldsymbol{\Theta}^{(0)}$ we get an improved guess

$$\boldsymbol{\Theta}^{(t+1)} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \sum_{\boldsymbol{Z} \in \{0,1\}^m} \mathbb{P}\Bigl[\boldsymbol{Z} \Big| \mathcal{D}, \boldsymbol{\Theta}^{(t)}\Bigr] \log\bigl(f(\mathcal{D}, \boldsymbol{Z}|\boldsymbol{\Theta})\bigr)$$

  - This is a general optimisation strategy that is regularly used when we have latent variables
  - It is known as **expectation maximisation** or the **EM-algorithm**
  - This looks very different to maximising the log-likelihood: it takes some effort to understand why this works
  - To understand this we note

$$f(\mathcal{D}, \boldsymbol{Z}|\boldsymbol{\Theta}) = f(\mathcal{D}|\boldsymbol{Z}, \boldsymbol{\Theta})\, \mathbb{P}[\boldsymbol{Z}|\boldsymbol{\Theta}]$$

    From which we can deduce

$$\log\bigl(f(\mathcal{D}|\boldsymbol{\Theta})\bigr) = \log\bigl(f(\mathcal{D}, \boldsymbol{Z}|\boldsymbol{\Theta})\bigr) - \log\bigl(\mathbb{P}[\boldsymbol{Z}|\boldsymbol{\Theta}]\bigr)$$

  - We now consider the probability distribution $\mathbb{P}\bigl[\boldsymbol{Z}\big|\mathcal{D}, \boldsymbol{\Theta}^{(t)}\bigr]$, that tells us the probability that $Z_i = 1$ given X_i and the parameters $\boldsymbol{\Theta}^{(t)}$
  - If we not take expectations of $\log\bigl(f(\mathcal{D}|\boldsymbol{\Theta})\bigr)$ give above with respect to this distribution then

$$\begin{aligned}\log\bigl(f(\mathcal{D}|\boldsymbol{\Theta})\bigr) &= \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{\Theta}^{(t)}}\bigl[\log\bigl(f(\mathcal{D}, \boldsymbol{Z}|\boldsymbol{\Theta})\bigr)\bigr] - \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{\Theta}^{(t)}}\bigl[\log\bigl(\mathbb{P}[\boldsymbol{Z}|\boldsymbol{\Theta}]\bigr)\bigr] \\ &= Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) + S(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})\end{aligned}$$

    * Note that the left-hand side does not involve the latent variables so when we take the expectation we get itself
    * The first term on the right-hand side is

$$Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{\Theta}^{(t)}}\bigl[\log\bigl(f(\mathcal{D}, \boldsymbol{Z}|\boldsymbol{\Theta})\bigr)\bigr] = \sum_{\boldsymbol{Z} \in \{0,1\}^m} \mathbb{P}\Bigl[\boldsymbol{Z} \Big| \mathcal{D}, \boldsymbol{\Theta}^{(t)}\Bigr] \log\bigl(f(\mathcal{D}|\boldsymbol{Z}, \boldsymbol{\Theta})\bigr)$$

    * This is the term we are optimising in *expectation maximisation*
    * The second term is

$$S(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) = -\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{\Theta}^{(t)}}\bigl[\log\bigl(\mathbb{P}[\boldsymbol{Z}|\boldsymbol{\Theta}]\bigr)\bigr] = -\sum_{\boldsymbol{Z} \in \{0,1\}^m} \mathbb{P}\Bigl[\boldsymbol{Z} \Big| \mathcal{D}, \boldsymbol{\Theta}^{(t)}\Bigr] \log\bigl(\mathbb{P}[\boldsymbol{Z}|\boldsymbol{\Theta}]\bigr)$$

  - Using the identity for the log-likelihood we can write the change in log-likelihood when we update our parameters

$$\begin{aligned}\Delta f &= \log\Bigl(f(\mathcal{D}|\boldsymbol{\Theta}^{(t+1)})\Bigr) - \log\Bigl(f(\mathcal{D}|\boldsymbol{\Theta}^{(t)})\Bigr) \\ &= Q(\boldsymbol{\Theta}^{(t+1)}|\boldsymbol{\Theta}^{(t)}) - Q(\boldsymbol{\Theta}^{(t)}|\boldsymbol{\Theta}^{(t)}) + S(\boldsymbol{\Theta}^{(t+1)}|\boldsymbol{\Theta}^{(t)}) - S(\boldsymbol{\Theta}^{(t)}|\boldsymbol{\Theta}^{(t)}) \\ &= Q(\boldsymbol{\Theta}^{(t+1)}|\boldsymbol{\Theta}^{(t)}) - Q(\boldsymbol{\Theta}^{(t)}|\boldsymbol{\Theta}^{(t)}) + \mathrm{KL}\Bigl(\mathbb{P}\Bigl[\boldsymbol{Z}|\boldsymbol{\Theta}^{(t)}\Bigr] \Big\| \mathbb{P}\Bigl[\boldsymbol{Z}|\boldsymbol{\Theta}^{(t+1)}\Bigr]\Bigr)\end{aligned}$$

* where

$$\mathrm{KL}\Big(\mathbb{P}\Big[\boldsymbol{Z}|\boldsymbol{\Theta}^{(t)}\Big]\Big\|\mathbb{P}\Big[\boldsymbol{Z}|\boldsymbol{\Theta}^{(t+1)}\Big]\Big) = S(\boldsymbol{\Theta}^{(t+1)}|\boldsymbol{\Theta}^{(t)}) - S(\boldsymbol{\Theta}^{(t)}|\boldsymbol{\Theta}^{(t)})$$
$$= -\sum_{\boldsymbol{Z}\in\{0,1\}^m} \mathbb{P}\Big[\boldsymbol{Z}\Big|\mathcal{D},\boldsymbol{\Theta}^{(t)}\Big]\ \log\left(\frac{\mathbb{P}\big[\boldsymbol{Z}|\boldsymbol{\Theta}^{(t+1)}\big]}{\mathbb{P}\big[\boldsymbol{Z}|\boldsymbol{\Theta}^{(t)}\big]}\right)$$

* We shown in a previous lecture that KL-divergences are non-negative

- Now in expectation maximisation we choose

$$\boldsymbol{\Theta}^{(t+1)} = \operatorname*{argmax}_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})$$

  which implies $Q(\boldsymbol{\Theta}^{(t+1)}|\boldsymbol{\Theta}^{(t)}) \geq Q(\boldsymbol{\Theta}^{(t)}|\boldsymbol{\Theta}^{(t)})$
- Thus $\Delta f \geq 0$
- This gives us a relative simple procedure we need to maximise

$$Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) = \sum_{\boldsymbol{Z}\in\{0,1\}^m} \mathbb{P}\Big[\boldsymbol{Z}\Big|\mathcal{D},\boldsymbol{\Theta}^{(t)}\Big]\ \log\big(f(\mathcal{D}|\boldsymbol{Z},\boldsymbol{\Theta})\big)$$

- Let us return to the problem of working out the half-life statistics of our two types of particles A and B
- Recall $f(\mathcal{D},\boldsymbol{Z}|\boldsymbol{\Theta}) = \prod_{i=1}^{m} f(X_i|Z_i,\boldsymbol{\Theta})\,\mathbb{P}[Z_i]$ where

$$f(X_i,Z_i|\boldsymbol{\Theta}) = p\,Z_i\,\mathcal{N}\big(X_i\big|\mu_A,\sigma_A^2\big) + (1-p)\,(1-Z_i)\,\mathcal{N}\big(X_i\big|\mu_B,\sigma_B^2\big)$$

- Let

$$p_i^{(t)} = \mathbb{P}\Big[Z_i=1\Big|X_i,\boldsymbol{\Theta}^{(t)}\Big] = \frac{p^{(t)}\,\mathcal{N}\Big(X_i|\mu_A^{(t)},\sigma_A^{2(t)}\Big)}{p^{(t)}\,\mathcal{N}\Big(X_i|\mu_A^{(t)},\sigma_A^{2(t)}\Big) + (1-p^{(t)})\,\mathcal{N}\Big(X_i|\mu_B^{(t)},\sigma_B^{2(t)}\Big)}$$
$$q_i^{(t)} = \mathbb{P}\Big[Z_i=0\Big|X_i,\boldsymbol{\Theta}^{(t)}\Big] = \frac{(1-p^{(t)})\,\mathcal{N}\Big(X_i|\mu_B^{(t)},\sigma_B^{2(t)}\Big)}{p^{(t)}\,\mathcal{N}\Big(X_i|\mu_A^{(t)},\sigma_A^{2(t)}\Big) + (1-p^{(t)})\,\mathcal{N}\Big(X_i|\mu_B^{(t)},\sigma_B^{2(t)}\Big)} = 1-p_i^{(t)}$$

- Then

$$Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) = \sum_{i=1}^{m} p_i^{(t)}\log\Big(p^{(t)}\,\mathcal{N}\big(X_i\big|\mu_A,\sigma_A^2\big)\Big) + q_i^{(t)}\log\Big((1-p^{(t)})\,\mathcal{N}\big(X_i\big|\mu_B,\sigma_B^2\big)\Big)$$
$$= \sum_{i=1}^{m} p_i^{(t)}\left(\log(p) - \frac{(X_i-\mu_A)^2}{2\sigma_A^2} - \frac{1}{2}\log\big(2\,\pi\,\sigma_A^2\big)\right)$$
$$+ q_i^{(t)}\left(\log(1-p) - \frac{(X_i-\mu_B)^2}{2\sigma_B^2} - \frac{1}{2}\log\big(2\,\pi\,\sigma_B^2\big)\right)$$

- To optimise this we just set the derivatives to 0
  * Optimising with respect to p

$$\frac{\partial Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})}{\partial p} = \frac{1}{p}\sum_{i=1}^{m} p_i^{(t)} - \frac{1}{1-p}\sum_{i=1}^{m} q_i^{(t)} = 0$$

  solving for p

$$p^{(t+1)} = \frac{\sum_{i=1}^{m} p_i^{(t)}}{\sum_{i=1}^{m}(p_i^{(t)}+q_i^{(t)})} = \frac{1}{m}\sum_{i=1}^{m} p_i^{(t)}$$

* Optimising with respect to μ_A

$$\frac{\partial Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})}{\partial \mu_A} = -\sum_{i=1}^{m} p_i^{(t)} \frac{X_i - \mu_A}{\sigma_A^2}$$

solving for μ_A (and performing a similar optimisation for μ_B)

$$\mu_A^{(t+1)} = \frac{\sum_{i=1}^{m} p_i^{(t)} X_i}{\sum_{i=1}^{m} p_i^{(t)}}, \qquad \mu_B^{(t+1)} = \frac{\sum_{i=1}^{m} q_i^{(t)} X_i}{\sum_{i=1}^{m} q_i^{(t)}}$$

* Putting in the optimal value for $\mu_A^{(t)}$ and optimising with respect to σ_A^2

$$\frac{\partial Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})}{\partial \sigma_A^2} = \frac{1}{2\,\sigma_A^4} \sum_{i=1}^{m} p_i^{(t)} (X_i - \mu_A^{(t)})^2 - \frac{1}{\sigma_A^2} \sum_{i=1}^{m} p_i^{(t)}$$

Solving for σ_A^2 (and performing a similar optimisation for σ_B^2)

$$\sigma_A^2 = \frac{\sum_{i=1}^{m} p_i^{(t)} (X_i - \mu_A^{(t)})^2}{\sum_{i=1}^{m} p_i^{(t)}}, \qquad \sigma_B^2 = \frac{\sum_{i=1}^{m} q_i^{(t)} (X_i - \mu_B^{(t)})^2}{\sum_{i=1}^{m} q_i^{(t)}}$$

- These are very natural update equations
  * we make an estimate, $p_i^{(t)}$ of the probability that observation X_i is a particle of type A or B base on our current parameters
  * we then update all our parameters based on these estimates
- We are guaranteed that our EM-algorithm always involves an improving step
- For the data set we showed earlier (which was a random sample of size 1000 generated using $p = 0.3$, $\mu_A = 4$, $\sigma_A = 0.8$, $\mu_B = 8$ and $\sigma_B = 2$ we get the results shown in figure 2

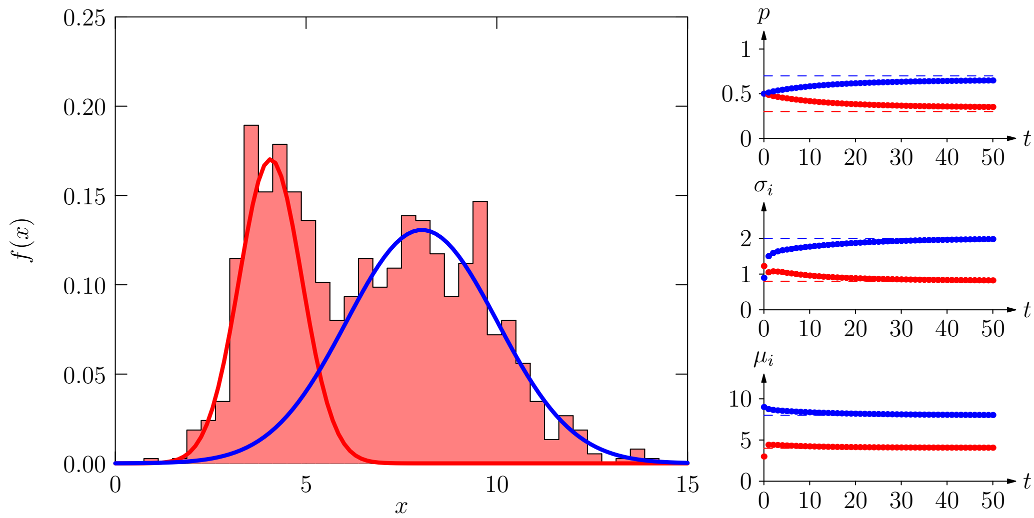

Figure 2: Example of EM algorithm to compute the statistics for the half-lives of our two particles

3 Exercises

**

4 Experiments

**