5MB20 - Exercises Part 1: Linear Gaussian Models and EM

course year 2010

Note: For some of these exercises you will be able to find solutions quickly on the internet. Try to resist this route to solving the problems. You will not be graded for these exercises and solutions will be made available through the class web page. **Your ability to solve these exercises without external help (apart from Sam Roweis' cheat sheets) provides an excellent indicator of your readiness to pass the exam**. Finally, the provided solutions come without guarantee and cannot be used as 'evidence' in case you choose to argue about your exam grade.

Ex. 1 — (lesson 2: Probability Theory). Box 1 contains 8 apples and 4 oranges. Box 2 contains 10 apples and 2 oranges. Boxes are chosen with equal probability.
(a) What is the probability of choosing an apple?
(b) If an apple is chosen, what is the probability that it came from box 1?

Ex. 2 — (lesson 2: Probability Theory). Derive the 'generalized sum rule',

$$p(\mathcal{A}+\mathcal{B}) = p(\mathcal{A}) + p(\mathcal{B}) - p(\mathcal{A},\mathcal{B})$$

from the 'elementary sum rule' $p(\mathcal{A}) + p(\bar{\mathcal{A}}) = 1$ and the product rule. Use the fact that $\mathcal{A}+\mathcal{B} = \overline{\bar{\mathcal{A}}\bar{\mathcal{B}}}$ (from Boolean logic).

Ex. 3 — (*DM03, ex.2.36*)[1]. (lesson 2: Probability Theory). The inhabitants of an island tell the truth one third of the time. They lie with probability 2/3. On an occasion, after one of them made a statement, you ask another 'was that statement true?' and he says 'yes'. What is the probability that the statement was indeed true?

Ex. 4 — (*DM03, ex.3.12*). (lesson 2: Probability Theory). A bag contains one ball, known to be either white or black. A white ball is put in, the bag is shaken, and a ball is drawn out, which proves to be white. What is now the chance of drawing a white ball? [Notice that the state of the bag, after the operations, is exactly identical to its state before.]

Ex. 5 — (lesson 2: Probability Theory). A dark bag contains five red balls and seven green ones.

[1] David Mackay, Information Theory, Inference, and Learning algorithms, Cambridge Press, 2003 (accessible at http://www.inference.org.uk/itila/book.html)

(a) What is the probability of drawing a red ball on the first draw? Balls are not returned to the bag after each draw.
(b) If you know that on the second draw the ball was a green one, what is now the probability of drawing a red ball on the first draw?

Ex. 6 — (lesson 2: Probability Theory). Consider two jointly distributed variables x and y, where x is an input and y is a response variable. As usual, the (joint) expectation is defined as $\mathbb{E}[g(x,y)] = \int_x \int_y g(x,y)p(x,y)\,\mathrm{d}x\,\mathrm{d}y$ and the *conditional expectation* as $\mathbb{E}[g(y)|x] = \int_y g(y)p(y|x)\,\mathrm{d}y$.
(a) Is $\mathbb{E}[y|x]$ a function of x only, of y only or is it a function of both x and y?
(b) Let's interpret $\hat{y} = \mathbb{E}[y|x]$ as an estimator for the outcomes y. Prove that for any function $f(x)$ of the input variables,

$$\mathbb{E}[yf^T(x)] = \mathbb{E}[\hat{y}f^T(x)]$$

(c) Prove that the conditional mean estimator $\mathbb{E}[y|x]$ minimizes the mean squared error loss function over all regression functions $f(x)$, i.e., prove that

$$\mathbb{E}\left[(y - f(x))^2\right] \geq \mathbb{E}\left[(y - \mathbb{E}[y|x])^2\right]$$

(d) Interpret this result.

Ex. 7 — (lesson 3: Bayesian Machine Learning). (a) Explain shortly the relation between machine learning and Bayes rule.
(b) How are Maximum a Posteriori (MAP) and Maximum Likelihood (ML) estimation related to Bayes rule and machine learning?

Ex. 8 — (lesson 5: density estimation).
We are given an IID data set $D = \{x_1, x_2, \ldots, x_N\}$, where $x_n \in \mathcal{R}^M$. Let's assume that the data were drawn from a multivariate Gaussian (MVG),

$$\begin{aligned} p(x_n|\theta) &= \mathcal{N}(x_n|\,\mu, \Sigma) \\ &= |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x_n - \mu)^T\Sigma^{-1}(x_n - \mu)\right\} \end{aligned}$$

(a) Derive the log-likelihood of the parameters for these data.

(b) Derive the maximum likelihood estimates for μ and Σ.

Ex. 9 — (lesson 5: density estimation).
Now we consider IID data $D = \{x_1, x_2, \ldots, x_N\}$ obtained from tossing a K-sided die. We use a *binary selection variable*

$$x_{nk} \equiv \begin{cases} 1 & \text{if } x_n \text{ lands on } k\text{th face} \\ 0 & \text{otherwise} \end{cases}$$

with probabilities $p(x_{nk} = 1) = \theta_k$.
(a) Write down the probability for the nth observation $p(x_n|\theta)$ and derive the log-likelihood $\ell(\theta; D) \equiv \log p(D|\,\theta)$.

(b) Derive the maximum likelihood estimate for θ.

Ex. 10 — (lesson 7: Generative Classification). You have a machine that measures property x, the 'orangeness' of liquids. You wish to discriminate between $C_1 =$ 'Fanta' and $C_2 =$ 'Orangina'. It is known that

$$p(x|C_1) = \begin{cases} 10 & 1.0 \leq x \leq 1.1 \\ 0 & \text{otherwise} \end{cases}$$

$$p(x|C_2) = \begin{cases} 200(x-1) & 1.0 \leq x \leq 1.1 \\ 0 & \text{otherwise} \end{cases}$$

The prior probabilities $p(C_1) = 0.6$ and $p(C_2) = 0.4$ are also known from experience.
(a) A 'Bayes Classifier' is given by

$$\text{Decide } C_1 \text{ if } p(C_1|x) > p(C_2|x); \text{ otherwise decide } C_2$$

Calculate the optimal Bayes classifier.
(b) The probability of making the wrong decision, given x, is

$$p(\text{error}|x) = \begin{cases} p(C_1|x) & \text{if we decide } C_2 \\ p(C_2|x) & \text{if we decide } C_1 \end{cases} \tag{1}$$

Compute the *total* error probability $p(\text{error})$ for the Bayes classifier in this example.

Ex. 11 — (lesson 7: Generative Classification and lesson 8: Discriminative Classification). Describe shortly in your own words the similarities and differences between the discriminative and generative approach to classification.

Ex. 12 — (Logistic regression). (lesson 8: Discriminative Classification). Given a data set $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $x_n \in \mathcal{R}^M$ and $y_n \in \{0, 1\}$. The probabilistic classification method known as *logistic regression* attempts to model these data as

$$p(y_n = 1|x_n) = \sigma(\theta^T x_n + b)$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the *logistic function*. Let's introduce shorthand notation $\mu_n = \sigma(\theta^T x_n + b)$. So, for every input x_n, we have a model output μ_n and an actual data output y_n.

(a) Express $p(y_n|x_n)$ as a Bernoulli distribution in terms of μ_n and y_n.
(b) If furthermore is given that the data set is IID, show that the log-likelihood is given by

$$\ell(\theta; D) = \sum_n \{y_n \log \mu_n + (1 - y_n) \log(1 - \mu_n)\}$$

(c) Prove that the derivative of the logistic function is given by

$$\sigma'(\xi) = \sigma(\xi) \cdot [1 - \sigma(\xi)]$$

(d) Show that the derivative of the log-likelihood is

$$\nabla_\theta \ell = \sum_{n=1}^{N} \big(y_n - \sigma(\theta^T x_n + b)\big)\, x_n$$

(e) Design a gradient-ascent algorithm for maximizing $\ell(\theta; D)$ with respect to θ.
(f) Interpret this result.

Ex. 13 — (lesson 9: Clustering). Consider a data set $D = \{x_1, x_2, \ldots, x_N\}$ and a MVG generative model for these data. The maximum likelihood estimate (MLE) of the mean value of this MVG distribution is given by

$$\hat{\mu} = \frac{1}{N} \sum_n x_n \tag{2}$$

In the lecture notes on linear generative classification, we derived the following expression for the MLE of the *class-conditional* mean,

$$\hat{\mu}_k = \frac{\sum_n t_{nk} x_n}{\sum_n t_{nk}} \tag{3}$$

(a) Explain this formula. What does t_{nk} represent? Relate this formula to the expression for the MLE of the mean for a MVG.

We then discussed MLE for a *clustering* problem and derived

$$\hat{\mu}_k^{(t)} = \frac{\sum_n r_n^{k(t)} x_n}{\sum_n r_n^{k(t)}} \tag{4}$$

(b) Again, explain this latter formula. What does r_n^k represent? Why the superscript (t)?

Let z_n^k be the usual binary selection variable, corresponding to the kth cluster.
(c) Express $r_n^{k(t)}$ as a conditional probability distribution (that involves both x_n and z_n).

Ex. 14 — (Factor Analysis). (lesson 11: Continuous Latent Variable models). Again we consider an observed data set $D = \{x_1, x_2, \ldots, x_N\}$ where $x_n \in \mathcal{R}^M$. This time we assume that the data were generated according to a model $x_n = \Lambda z_n + v_n$ where $z_n \in \mathcal{R}^K$, $K < M$. The samples z_n are IID drawn from a distribution $\mathcal{N}(0, I)$ and $v_n \sim \mathcal{N}(0, \Psi)$. Furthermore, we assume that z_n and v_n are uncorrelated, i.e., $\epsilon[z_n v_n^T] = 0$.

(a) Write this model in terms of $p(x_n|z_n)$ and a prior $p(z_n)$.

Note that $p(x_n)$ is a Gaussian distribution since products and marginals of Gaussian distributions yield again Gaussian distributions.
(b) Given that $p(x_n)$ is Gaussian, proof that

$$p(x_n) = \mathcal{N}(0, \Lambda\Lambda^T + \Psi)$$

(hint: workout the expectation and variance of x_n).
(c) Why is this model not very interesting if the only constraint for Ψ is that it's a symmetric positive definite matrix? What's so interesting about this model if Ψ is constrained to be a diagonal matrix?

In the E-step of an EM-algorithm for estimating Ψ, we compute the posterior distribution of hidden factors z_n, given the observed data, as

$$\begin{aligned} q_n^{(t+1)}(z) &= p(z|x_n, \theta^{(t)}) = \mathcal{N}(z|\, m_n, V) \\ V &= (I + \Lambda^T \Psi^{-1} \Lambda)^{-1} \\ m_n &= V \Lambda^T \Psi^{-1} x_n \end{aligned}$$

In the M-step, we then maximize the free energy function w.r.t Ψ and obtain

$$\hat{\Psi}^{(t+1)} = \text{diag} \left\{ \Lambda^{(t+1)} V (\Lambda^{(t+1)})^T + \frac{1}{N} \sum_n (x_n - \Lambda m_n)(x_n - \Lambda m_n)^T \right\}$$

(d) Use the expression for $\hat{\Psi}^{(t+1)}$ to explain differences (and similarities) between this model and the linear regression model.

Ex. 15 — (Temporal Models) (Lesson 13: Dynamic latent variable models). What is the 1st-order Markov assumption? Derive the joint probability distribution $p(x_{1:T}, z_{1:T})$ from transition and observation models ($p(z_t|z_{t-1})$ and $p(x_t|z_t)$). What is a HMM? What is a Kalman Filter? What is a Linear Dynamizal System (LDS)? How does the Kalman Filter relate to the LDS? And to FA? Explain the popularity of Kalman filtering and HMMs? How relates a HMM to a GMM? What's the significance of the α and β recursions?

Ex. 16 — Work through previous exams!!