Contents

1 Exercises Mackay: Information Theory, Inference, and Learning Algorithms **2**
- 1.1 Chapter 2 . . . 2
- 1.2 Chapter 3 . . . 3
- 1.3 Chapter 27 . . . 3
- 1.4 Chapter 28 . . . 4
- 1.5 Chapter 29 . . . 5
- 1.6 Chapter 31 . . . 5
- 1.7 Chapter 34 . . . 5

2 Other exercises **6**
- 2.1 Perceptron . . . 6
- 2.2 Graphical models . . . 7
- 2.3 Mixture models and EM . . . 8

1 Exercises Mackay: Information Theory, Inference, and Learning Algorithms

1.1 Chapter 2

1. (A twist on exercise 2.6). Consider 11 urns $u = 0, \ldots, 10$ each with 10 balls. Urn u has u black balls and $10-u$ white balls. Select one urn at random, and draw N times with replacement from that urn. Suppose that the outcome after $N = 10$ draws is that the number of black balls that have been drawn is *even*. What is the probability that urn u was selected?

2. (Inference on the parameters of a Gaussian distribution). The Gaussian distribution is the probability density of a continuous variable $x \in \mathbb{R}$ and has the form

$$p(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

   μ is the peak of the Gaussian distribution and also the expected value of x: $\mu = \mathbb{E}x = \int dx x p(x|\mu, \sigma)$, where $\mathbb{E}$ denotes expected value. σ^2 is the variance of x: $\sigma^2 = \mathbb{E}(x-\mu)^2 = \int dx (x-\mu)^2 p(x|\mu, \sigma)$.

   Suppose that we are told that data are drawn from an oracle that is a Gaussian distribution with $\sigma = 1$ and we believe that μ can have any value with equal probability.

   (a) The oracle produces one data point with value x. What can we infer about μ?

   (b) The oracle produces a set of N data points with values $x_1, \ldots, x_N$. What can we infer about μ? Hint: use the fact that the data points are produced independently. What is the probability that the data set is drawn from the oracle? Show that the result is Gaussian in μ with mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and variance $\frac{1}{N}$.

3. (Maximum entropy distributions). Consider the n-dimensional multivariate Gaussian distribution

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

   with $x, \mu \in \mathbb{R}^n$ and Σ a positive definite $n \times n$ matrix.

   (a) Show that this distribution is an exponential family distribution.

   (b) The multivariate Gaussian distribution is the a maximum entropy solution. What are the constraints?

4. (Variational and reverse KL approximation). Consider a two-dimensional Gaussian distribution

$$p(z_1, z_2) \propto \exp\left(-\frac{1}{2}(az_1^2 + az_2^2 + 2bz_1 z_2)\right)$$

   p has mean zero: $\mathbb{E}_p z_1 = \mathbb{E}_p z_2 = 0$ and covariance $\mathbb{E}_p z_i z_j = \left(\Lambda^{-1}\right)_{ij}$ with $\Lambda = \begin{pmatrix} a & b \\ b & a \end{pmatrix}$. We wish to approximate p by a simpler spherical Gaussian distribution

$$q(z_1, z_2) = q_1(z_1) q_2(z_2) \qquad q_i(z_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-\frac{1}{2\sigma_i^2} z_i^2)$$

(Since p has mean zero, we already know that we should take q also mean zero). First, we compute the variational solution in the following steps.

(a) Show that $\int dz_1 q_1(z_1) \log q_1(z_1) = -\log \sqrt{2\pi\sigma_1^2} - \frac{1}{2}$ and similar for $q_2(z_2)$.

(b) Use this result to show that

$$KL(q|p) \quad = \quad -1 - \log \sqrt{2\pi\sigma_1^2} - \log \sqrt{2\pi\sigma_2^2} + \frac{1}{2}a(\sigma_1^2 + \sigma_2^2)$$

(c) Show that the variational solution is given by $q_i(z_i)$ with $\sigma_i^2 = \frac{1}{a}$.

(d) Now we compute the reverse solution. Show that reverse variational solution is given by $q_i(z_i)$ with $\sigma_i^2 = \mathbb{E}_p z_i^2, i = 1, 2$.

(e) Show that $\mathbb{E}_p z_1^2 = \mathbb{E}_p z_2^2 = \frac{a}{a^2 - b^2}$. Thus the variance of the variational solution is less that the (correct) variance of the reverse variational solution:

$$\frac{1}{a} < \frac{a}{a^2 - b^2}$$

unless $b = 0$, in which case p itself is a factorized distribution.

1.2 Chapter 3

1. Consider the bent coin model comparison example of Mackay section 3.2-3 with $N = 2$, where you take as model $\mathcal{H}_0$ that the coin is fair with probability of 'head' $f = 0.5$.

   (a) Compute the posterior probability of the two models $H_{0,1}$ for $N_H = 0, 1, 2$.

   (b) You will find that for $N_H = 0, 2$ model H_1 is more likely and for $N_H = 1$ model H_0 is more likely. Explain these results.

1.3 Chapter 27

1. (Adapted from Mackay 27.1). A photon counter is pointed at a remote star for one minute, in order to infer the rate of photons arriving at the counter per minute, λ. Assuming the number of photons collected r has a Poisson distribution with mean λ

$$p(r|\lambda) = \exp(-\lambda)\frac{\lambda^r}{r!}$$

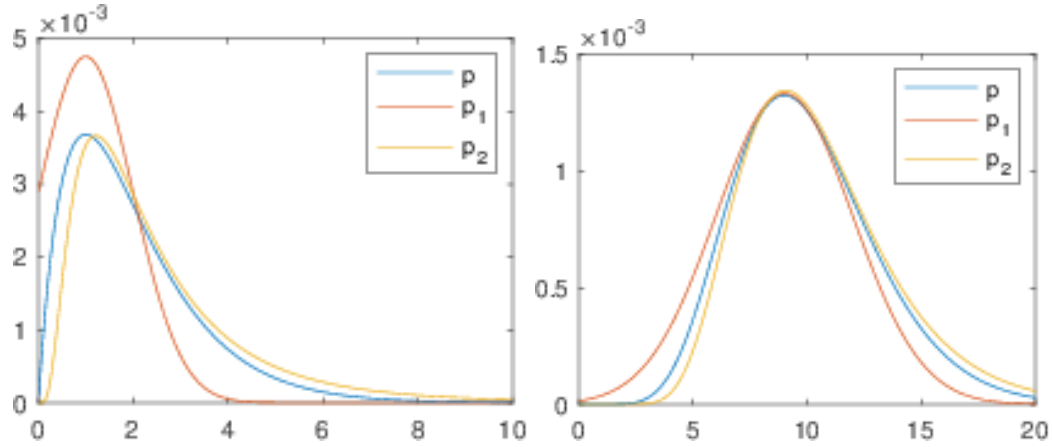
and assuming the improper prior $p(\lambda) = \frac{1}{\lambda}$.

   (a) compute the Laplace approximations $p_1(\lambda)$ to the posterior distribution

   (b) Consider the transformation to new coordinate $y = \log \lambda$. show that the prior distribution over y transforms to $p(y) = 1$. Hint: When changing variables, the distribution changes as $p(\lambda)d\lambda = p(y)dy$.

   (c) Transform the posterior distribution to the new variable y and compute the Laplace approximation $p_2(y)$.

(d) Transform $p_2(y)$ back to $p_2(\lambda)$. Plot the posterior $p(\lambda|r)$ and its two approximations $p_1(\lambda), p_2(\lambda)$ for $r = 2$ and $r = 10$ versus λ. Hint: Devide all distributions by their sum to get comparable magnitudes not not have to worry about explicit normalization. It should look like this



Which is better and why?

1.4 Chapter 28

1. (Variant on Mackay Ex. 28.2) Data points $D = (x_i, t_i), i = 1, \ldots, N$ are believed to come from a straight line with noise. The experimenter chooses x_i and assumes that t_i is Gaussian distributed about

$$w_0 + w_1 x_i$$

with variance σ^2. According to model $\mathcal{H}_1$, the straight line is horizontal, so $w_1 = 0$. According to model $\mathcal{H}_2$, w_1 is a parameter with prior distribution a Normal Gaussian distribution $\mathcal{N}(w_1|0, 1)$. Both models assign a prior distribution $\mathcal{N}(w_0|0, 1)$ to w_0. Given the data, what is the evidence for each model? You may assume that the input data x_i are 'centered" such that $\sum_{i=1}^{N} x_i = 0$. Proceed in the following way.

   (a) Write an expression for $p(D|w_0, w_1)$. Note, that the result is a product of a function of w_0 and a function of w_1.

   (b) You can now compute the odds ratio $p(D|\mathcal{H}_2)/p(D|\mathcal{H}_1)$. Show the result as a function of the input variance $\left\langle x^2 \right\rangle = \frac{1}{N}\sum_i x_i^2$, the input-output correlation $\langle xt \rangle = \frac{1}{N}\sum_i x_i t_i$ and N. Hint: make use of the fact that the integrals over w_0, w_1 factorize. As a result many terms cancel in the odds ratio $p(D|\mathcal{H}_2)/p(D|\mathcal{H}_1)$ and you dont have to compute some of the integrals.

   (c) Show that in the limit of large N the simpler model H_1 is preferred when the correlations $\langle xt \rangle^2 \lesssim \frac{\log N}{N}$ and the complex model is preferred otherwise. Use $\sigma^2 = \langle x \rangle^2 = 1$.

2. (Variant on Mackay Ex. 28.3) A k sided die is rolled n times and the outcomes are $D = (n_1, \ldots, n_k)$ where n_i is the number of times the die lands with side i up, $i = 1, \ldots, k$ and $n = \sum_{i=1}^{k} n_i$.

   (a) Compute the probability of the observed data assuming that the die is perfectly fair ($\mathcal{H}_0$). Hint: the data $D = (n_1, \ldots, n_k)$ specifies that we get n_1 times the outcome 1, etc. The

data does not specify the order, just the frequencies. So the probability of the data is a combinatorial factor times the probability of a specific order. The combinatorial factor for a partition of $n_1, \ldots, n_k$ of n with $\sum_{i=1}^{k} n_i = n$ is

$$\frac{n!}{n_1! \ldots n_k!}$$

(b) Compute the probability of the data under the alternative hypothesis $\mathcal{H}_1$ that the die has an unknown probability p_i for outcome i and that p_i is uniformly distributed. Use the fact that the probability density function of the Dirichlet distribution with parameters $\vec{\alpha}$ is $p(\vec{p}|\vec{\alpha}) = \frac{1}{B(\vec{\alpha})} \prod_{i=1}^{k} p_i^{\alpha_i - 1}$. The pdf lives on the simplex $0 \leq p_i \leq 1$ with $\sum_{i=1}^{k} p_i = 1$. The normalization is given by the integral

$$B(\vec{\alpha}) = \int_{\text{simplex}} dp_1 \ldots dp_k \prod_{i=1}^{k} p_i^{\alpha_i - 1} = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)}$$

(c) Consider $k = 6, n = 30$ and $(n_1, n_2, n_3, n_4, n_5, n_6) = (3, 3, 2, 2, 9, 11)$. Compute the posterior probability of the models $\mathcal{H}_{0,1}$ assuming equal priors. Show that $\mathcal{H}_1$ is more likely. Instead, when $n_1, n_2, n_3, n_4, n_5, n_6) = (5, 5, 5, 5, 5, 5)$ show that $\mathcal{H}_0$ is more likely.

1.5 Chapter 29

1. Show that $\hat{\Phi}$ as given in eq. 29.6 is an unbiased estimator of Φ as given by eq. 29.3 and that the variance in $\hat{\Phi}$ decreases with R.

1.6 Chapter 31

1. Show, that the Ising model Eq. 31.1 on the square lattice with ferromagnetic interactions (all $J = 1$ and $H = 0$) has energy per spin -2.

2. Show analytically by dividing the spins into two groups, that the infinite lattice with antiferromagnetic interactions (all $J = -1$ and $H = 0$) can be transformed into an equivalent ferromagnetic model.

3. Use the result of MacKay exercise 31.1c to explain the slope of F vs. T in Fig. 31.15.

1.7 Chapter 34

1. Derive Eq. 34.11.

2. Derive Eq. 34.12

3. Derive the ICA algorithm from the likelihood 34.9.

2 Other exercises

2.1 Perceptron

1. In this exercise we look at some special cases of the perceptron capacity

$$C(P,N) = 2\sum_{i=0}^{N-1}\left(\begin{array}{c} P-1 \\ i \end{array}\right)$$

to better understand the behaviour of C.

   (a) Show that all problems with $P \leq N$ are linearly separable.

   (b) Show that exactly half of the problems with $P = 2N$ are linearly separable.

Hint: In the formula, the convention is $\left(\begin{array}{c} n \\ 0 \end{array}\right) = 1; \left(\begin{array}{c} n \\ k \end{array}\right) = 0$ when $n < k$ assumed. Other hint: The formula $(1+\alpha)^n = \sum_{i=0}^{n}\binom{n}{i}\alpha^i$ may come in handy.

2. In this exercise we numerically check the formula $C(N,P)$ for the number of linearly separable problems.

   (a) Write a computer program that implements the perceptron learning rule. Take as data P random input vectors of dimension N with binary components. Take as outputs random assignments ± 1.

   (b) Take $N = 50$. Test empirically for individual problems that when $P < 2N$ the rule converges almost always and for $P > 2N$ the rule converges almost never.

   (c) Reconstruct the curve $C(P,N)$ for $N = 50$ as a function of P in the following way. For each P construct a number (nruns) of learning problems randomly and compute 1) the fraction of these problems for which the perceptron learning rule converges, 2) the mean and std of the classification error on the training set and 3) the mean and std of the number of iterations until convergence.

Suggestions: Use $P = 10, 20, 30, \ldots, 120$; Take $nruns = 100$. Decide that the algorithm does not converge when 1000 iterations has been reached.

3. The number of linearly separable problems of P patterns in N dimensions is given by $C(N,P)$. We know that $C(N,P) = 2^P$ when $P \leq N$. When $P > N$ we can use the bound

$$C(N,P) \leq \left(\frac{eP}{N}\right)^N$$

Compute numerically $C(N,P)$ and its bound for $N = 50$ and for $P = 1$ to $P = 200$.

4. The generalization bound is quite conservative. In this excercise we will verify this numerically for the perceptron. We define

$$\delta \;=\; 4m(2P)\exp\left(-\frac{\epsilon^2 P}{8}\right)$$

and put it for instance to $\delta = 0.01$. We can then ask what the error ϵ is for given N and P. We can compare this error with the generalization error that we find by numerical simulation.

In particular, suppose that data is generated from a so-called teacher perceptron which is specified by an N dimensional weight vector w^{teacher}. The input data are P binary vectors, each of dimension N. So we can define the input data as a matrix ξ of size $N \times P$. We generate a training set by defining output labels $y_j = \text{sign}\left(\sum_{i=1}^{N} \xi_{ij} w_i^{\text{teacher}}\right)$.

The training data are used to train another perceptron (the so-called student perceptron). By construction the problem is linearly separable and therefore the perceptron learning rule will always converge and the solution will perfectly separate the two classes. Thus, in terms of the generalization bound, the student solution implements a function f with $g_P(f) = 1$. The probability that the generalization performance $g(f)$ of this solution is larger than $1 - \epsilon$ is given by the generalization bound.

We can get a numerical estimate of the generalization error, by generating a separate test set of P_{test} patterns with labels again computed from the teacher perceptron. The generalization error is the fraction of test patterns that are incorrectly classified by the student perceptron solution.

The student solution f is not unique. Starting with a different initial weight vector, a different converged solution f is obtained. In order to get a reliable numerical estimate of the generalization error, we should run the perceptron learning rule many times with different initial weight vectors and compute the average generalization error.

(a) Using the formula for δ above compute an expression of ϵ in terms of N and P and $\delta = 0.01$. Approximate $m(P) = C(N, P)$ by its bound as given in exercise 3. Compute numerically for $N = 10$ the dependence of ϵ on P. Compute the number of patterns P to ensure that $\epsilon \approx 0.1$. Repeat this for $N = 20, 30, 40, 50$. Note, that the required number of patterns scales linearly with N.

(b) Estimate the generalization error for the teacher student perceptron learning scenario as described above. Detais:

i. Generate input training data ξ of size $N \times P$ with ξ_{ij} binary ± 1.

ii. Define a random (but fixed) teacher vector w^{teacher}: `w_0=randn(1,n)`.

iii. Compute the teacher labels y_j as defined above

iv. Generate in the same way a test set ξ_{test} of size $N \times P_{\text{test}}$ with $P_{\text{test}} = 10.000$ and teacher labels.

v. Compute `n_learning_runs=100` perceptron solutions by training on the training set with P samples with different initial weight vectors w. After convergence, the training error should be zero ($g_P(f) = 1$) but the solutions are different. Compute for each solution the generalization error on the test set ϵ. Use $N = 10$ and $P = 10, 50, 100, 500, 1000$.

vi. Make a table where you compare your numerical estimates for ϵ with those given by the generalization bound ($\delta = 0.01$).

2.2 Graphical models

1. Consider the problem of a variable y whose value can depend on n possible causes $x_i, i = 1, \ldots, n$. For instance, in medical diagnosis, y may be a symptom that is observed in a patient ('fever') and $x_i = 0, 1$ may be one of the diseases that is causing the fever. When the causes are marginally independent the full model is $p(x_1, x_2, \ldots, x_n, y) = p(x_1)p(x_2)\ldots p(x_n)p(y|x_1, x_2, \ldots, x_n)$.

(a) When y is observed, we can infer the probability that it is caused by disease x_i by computing the probability $p(x_i|y), i = 1, \ldots, n$. Exercise: give an expression for $p(x_i|y)$.

(b) Consider now the case that we have independent information that in fact the patient has disease 1 ($x_1 = 1$). Exercise: give an expression for $p(x_i|y, x_1 = 1)$ for $i \neq 1$.

The phenomenon of explaining away is that $p(x_i = 1|y, x_1 = 1) < p(x_i = 1|y)$: the certainty that disease 1 is present ($x_1 = 1$), partly explains the observed fever y and makes the other causes x_i less likely. This is not so easy to see in the tabular case (binary x), but easy to see in the Gaussian case. We study this in detail for a Gaussian directed model with two causes and one effect with $x_{1,2}$ two standard Normal variables $p(x_i) = \mathcal{N}(x_i|0, 1), i = 1, 2$ and y depends on $x_{1,2}$ as $p(y|x_1, x_2) = \mathcal{N}(y|\gamma(x_1 + x_2), \sigma^2)$.

We will compute the correlation between x_1 and x_2 in the condition distribution $p(x_1, x_2|y)$. Proceed in the following way

(a) We first compute the joint distribution $p(x_1, x_2, y) = p(x_1)p(x_2)p(y|x_1, x_2)$ which is a Gaussian distribution because it is a product of Gaussian distributions. Therefore, it is fully specified by its mean and covariance matrix. Hint: Note, that we can express the conditional distribution of y given x_1, x_2 equivalently by saying that

$$y = \gamma(x_1 + x_2) + \xi$$

with ξ a Gaussian random variable with mean zero and variance $\mathbb{E}\xi^2 = \sigma^2$.

(b) Compute the conditional distribution $p(x_1, x_2|y)$. Hint: use the general expression for the conditional distribution of a Gaussian, as given at the end of the lecture slides

(c) With Σ the covariance matrix between x_1, x_2 in the posterior distribution $p(x_1, x_2|y)$, show that the correlation coefficient $\rho = \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}}$ is given by

$$\rho = \frac{-\gamma^2}{\gamma^2 + \sigma^2}$$

While a priori x_1, x_2 are independent (correlation is zero), when y is observed these variables become anti-correlated.

This result explains the explaining away in the tabular binary case as well, if we are willing to accept that in the binary case something similar will happen as in the Gaussian case, and thus that x_1 and x_2 are anti-correlated when y is observed. This means that observing $x_1 = 1$ increases the probability that $x_2 = 0$ making the cause x_2 less likely, and vice versa.

2.3 Mixture models and EM

1. Write a clustering algorithm based on the multinomial mixture model and apply it to the MNIST data.

2. Consider the one dimensional Gaussian mixture model

$$p(x, k) = \pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x - a_k)^2}{2\sigma_k^2}\right)$$

with observable data $x^\mu, \mu = 1, \ldots, N$ and discrete latent variable $k = 1, \ldots, K$. Derive an EM algorithm to estimate the parameters $\pi_k, a_k, \sigma_k^2, k = 1, \ldots, K$ from the data. Proceed with the following steps.

(a) Give an expression for the responsabilities r_k^μ that result from the E step

(b) Give an expression for the variational bound $Q(\theta, q^*)$ in terms of the responsabilities.

(c) Show that the M-step can be solved in close form and yields new values of $\pi_k, a_k, \sigma_k^2, k = 1, \ldots, K$ in terms of the responsabilities and the data. Check that your final result agrees with the multi-dimensional case presented in the slides.