My Solutions for Exercises of Deep Learning Fundamentals by Bishop

Andres Espinosa

May 15, 2024

Contents

1 The Deep Learning Revolution **4**
1.1 No Exercises . . . 4

2 Probabilities **5**
2.1 Exercise 2.2 . . . 5
2.2 Exercise 2.4 . . . 5
2.3 Exercise 2.6 . . . 6
2.4 Exercise 2.7 . . . 6
2.5 Exercise 2.8 . . . 6
2.6 Exercise 2.9 . . . 6
2.7 Exercise 2.10 . . . 6
2.8 Exercise 2.11 . . . 7
2.9 Exercise 2.12 . . . 7
2.10 Exercise 2.13 . . . 7
2.11 Exercise 2.14 . . . 7
2.12 Exercise 2.15 . . . 8
2.13 Exercise 2.16 . . . 8
2.14 Exercise 2.17 . . . 8
2.15 Exercise 2.18 . . . 9
2.16 Exercise 2.19 . . . 9
2.17 Exercise 2.20 . . . 9
2.18 Exercise 2.21 . . . 10
2.19 Exercise 2.22 . . . 10
2.20 Exercise 2.23 . . . 10
2.21 Exercise 2.24 . . . 10
2.22 Exercise 2.25 . . . 10
2.23 Exercise 2.26 . . . 10
2.24 Exercise 2.27 . . . 10
2.25 Exercise 2.28 . . . 10
2.26 Exercise 2.29 . . . 10
2.27 Exercise 2.30 . . . 11
2.28 Exercise 2.31 . . . 11
2.29 Exercise 2.32 . . . 11
2.30 Exercise 2.33 . . . 11

2.31 Exercise 2.34 . . . 11
2.32 Exercise 2.35 . . . 11
2.33 Exercise 2.36 . . . 11
2.34 Exercise 2.37 . . . 11
2.35 Exercise 2.38 . . . 11
2.36 Exercise 2.39 . . . 11
2.37 Exercise 2.40 . . . 11
2.38 Exercise 2.41 . . . 11

3 Standard Distributions 12
3.1 Exercise 3.1 . . . 12
3.2 Exercise 3.2 . . . 12
3.3 Exercise 3.3 . . . 13
3.4 Exercise 3.4 . . . 14
3.5 Exercise 3.5 . . . 14
3.6 Exercise 3.6 . . . 14
3.7 Exercise 3.7 . . . 14
3.8 Exercise 3.8 . . . 15
3.9 Exercise 3.9 . . . 15
3.10 Exercise 3.10 . . . 15
3.11 Exercise 3.11 . . . 15
3.12 Exercise 3.12 . . . 15
3.13 Exercise 3.13 . . . 15
3.14 Exercise 3.14 . . . 16
3.15 Exercise 3.15 . . . 16
3.16 Exercise 3.16 . . . 16
3.17 Exercise 3.17 . . . 16
3.18 Exercise 3.18 . . . 16
3.19 Exercise 3.19 . . . 16
3.20 Exercise 3.20 . . . 16
3.21 Exercise 3.21 . . . 16
3.22 Exercise 3.22 . . . 16
3.23 Exercise 3.23 . . . 16
3.24 Exercise 3.24 . . . 16
3.25 Exercise 3.25 . . . 16
3.26 Exercise 3.26 . . . 16
3.27 Exercise 3.27 . . . 17
3.28 Exercise 3.28 . . . 17
3.29 Exercise 3.29 . . . 17
3.30 Exercise 3.30 . . . 17
3.31 Exercise 3.31 . . . 17
3.32 Exercise 3.32 . . . 17
3.33 Exercise 3.33 . . . 17
3.34 Exercise 3.34 . . . 17
3.35 Exercise 3.35 . . . 17
3.36 Exercise 3.36 . . . 17
3.37 Exercise 3.37 . . . 17
3.38 Exercise 3.38 . . . 18

4 Single-layer Networks: Regression **19**
4.1 Exercise 4.1 19
4.2 Exercise 4.2 19
4.3 Exercise 4.3 19
4.4 Exercise 4.4 20
4.5 Exercise 4.5 20
4.6 Exercise 4.6 21
4.7 Exercise 4.7 21
4.8 Exercise 4.8 21
4.9 Exercise 4.9 22
4.10 Exercise 4.10 22
4.11 Exercise 4.11 22
4.12 Exercise 4.12 22

1 The Deep Learning Revolution

1.1 No Exercises

2 Probabilities

Exercise 2.1

Bayes rule

$$P[C=1|T=1] = \frac{P[T=1|C=1]*P[C=1]}{P[T=1|C=1]*P[C=1]+P[T=1|C=0]*P[C=0]} \tag{1}$$

$$P[C=1|T=1] = \frac{0.90*0.001}{0.90*0.001+0.03*0.999} = 0.0292 \tag{2}$$

Given that the test result was positive, there is a 2.92% chance that you have cancer.

2.1 Exercise 2.2

Not attempted

Exercise 2.3

$$p(\mathbf{y}) = \int p_{\mathbf{u},\mathbf{v}}\left(\mathbf{u}, \mathbf{y}-\mathbf{u}\right) d\mathbf{u} \tag{3}$$

$$= \int p_{\mathbf{u}}(\mathbf{u}) p_{\mathbf{v}}(\mathbf{y}-\mathbf{u}) d\mathbf{u} \tag{4}$$

2.2 Exercise 2.4

Not attempted

Exercise 2.5

Exponential:

$$p(x|\lambda) = \lambda e^{-\lambda x} \tag{5}$$

Laplace:

$$p(x|\mu,\gamma) = \frac{1}{2\gamma} e^{-\frac{|x-\mu|}{\gamma}} \tag{6}$$

Verifying that the exponential distribution is normalized:

$$p(x|\lambda) = \lambda e^{-\lambda x} \tag{7}$$

$$\int_0^\infty \lambda e^{-\lambda x} = -e^{-\lambda x}\Big|_0^\infty = \frac{1}{e^\infty} + \frac{1}{e^0} \tag{8}$$

$$= 1 \tag{9}$$

Verifying the laplace distribution:

$$p(x|\mu,\gamma) = \frac{1}{2\gamma} e^{-\frac{|x-\mu|}{\gamma}} \tag{10}$$

$$\begin{cases} \frac{1}{2\gamma} e^{-\frac{x-\mu}{\gamma}} & \text{if } x \geq \mu \\ \frac{1}{2\gamma} e^{-\frac{-x+\mu}{\gamma}} & \text{if } x < \mu \end{cases} \tag{11}$$

$$\int_{\mu}^{\infty} \frac{1}{2\gamma} e^{-\frac{x-\mu}{\gamma}} = \frac{1}{2} e^{-\frac{x-\mu}{\gamma}} \Big|_{\mu}^{\infty} = -\frac{1}{2}(e^{-\infty} - e^{0}) \tag{12}$$

$$= \frac{1}{2} \tag{13}$$

$$\int_{-\infty}^{\mu} \frac{1}{2\gamma} e^{-\frac{-x+\mu}{\gamma}} = \frac{1}{2} e^{-\frac{-x+\mu}{\gamma}} \Big|_{-\infty}^{\mu} = \frac{1}{2}(e^{0} - e^{-\infty}) \tag{14}$$

$$= \frac{1}{2} \tag{15}$$

$$\frac{1}{2} + \frac{1}{2} = 1 \tag{16}$$

2.3 Exercise 2.6

Not attempted

2.4 Exercise 2.7

$$P(x|D) = \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n) \tag{17}$$

$$E[f] = \int p(x) f(x) dx \tag{18}$$

$$\text{Substituting:} \tag{19}$$

$$E[f] = \int \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n) f(x) dx \tag{20}$$

$$E[f] = \frac{1}{N} \sum_{n=1}^{N} \int_{x_n-\varepsilon}^{x_n+\varepsilon} \delta(x - x_n) f(x) dx \tag{21}$$

$$E[f] = \frac{1}{N} \sum_{n=1}^{N} f(x_n) \int_{x_n-\varepsilon}^{x_n+\varepsilon} \delta(x - x_n) dx \tag{22}$$

$$E[f] = \frac{1}{N} \sum_{n=1}^{N} f(x_n) \tag{23}$$

2.5 Exercise 2.8

Not attempted

2.6 Exercise 2.9

$$cov[X, y] = E_{x,y}[xy] - E[x]E[y] \tag{24}$$

If x and y are independent, the joint distribution is equal to the product of the marginals. $p(x, y) = p(x)p(y)$. If $E_{x,y}[xy] = E[x]E[y]$, then the covariance will be zero.

2.7 Exercise 2.10

Not attempted

2.8 Exercise 2.11

Proving $E[x] = E_y[E_x[x|y]]$:

$$E_x[x|y] = \int p(x|y)xdx \tag{25}$$

$$\text{Substituting:} \tag{26}$$

$$E[x] = E_y[\int p(x|y)xdx] \tag{27}$$

$$E[x] = \int E_y[p(x|y)]xdx \tag{28}$$

$$E[x] = \int\int p(x|y)xp(y)dxdy \tag{29}$$

$$E[x] = \int\int \frac{p(x,y)}{p(y)}xp(y)dxdy \tag{30}$$

$$E[x] = \int\int p(x,y)xdxdy \tag{31}$$

$$E[x] = E[x] \tag{32}$$

2.9 Exercise 2.12

Not attempted

2.10 Exercise 2.13

$$E[x] = \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{1}{2\sigma^2}(x-\mu)^2}dx \tag{33}$$

$$\text{Change of variables } z = \frac{x-\mu}{\sigma}, \sigma dz = dx \tag{34}$$

$$E[x] = \int_{-\infty}^{\infty} \sigma\frac{\sigma z+\mu}{\sqrt{2\pi\sigma^2}}e^{\frac{1}{2}z^2}dz \tag{35}$$

$$E[x] = \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty} ze^{\frac{1}{2}z^2} + \frac{\mu}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{1}{2}z^2} \tag{36}$$

$$E[x] = \frac{\sigma}{\sqrt{2\pi}} * 0 + \frac{\mu}{\sqrt{2\pi}} * \sqrt{2\pi} \tag{37}$$

$$E[x] = \mu \tag{38}$$

2.11 Exercise 2.14

Not attempted

2.12 Exercise 2.15

Solving for μ_{ml}:

$$\log p(x|\mu, \sigma^2) = \frac{-1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log 2\pi \tag{39}$$

$$\frac{d}{d\mu} logp(x|\mu, \sigma^2) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} 2(x_n - \mu) \tag{40}$$

$$0 = \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu) \tag{41}$$

$$0 = \sum_{n=1}^{N} x_n - \sum_{n=1}^{N} \mu \tag{42}$$

$$N\mu = \sum_{n=1}^{N} x_n \tag{43}$$

$$\mu_{ml} = \frac{1}{n} \sum_{n=1}^{N} x_n \tag{44}$$

Solving for σ_{ml}:

$$\log p(x|\mu, \sigma^2) = \frac{-1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \log \sigma^2 - \frac{N}{2} \log 2\pi \tag{45}$$

$$\frac{d}{d\sigma^2} logp(x|\mu, \sigma^2) = \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2\sigma^2} \tag{46}$$

$$\frac{N}{2\sigma^2} = \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu)^2 \tag{47}$$

$$\sigma_{ml}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ml})^2 \tag{48}$$

2.13 Exercise 2.16

not attempted

2.14 Exercise 2.17

Finding expectation of $\hat{\sigma}^2$

$$E[\hat{\sigma}^2] = E[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2] \tag{49}$$

$$= \frac{1}{N}\sum_{n=1}^{N} E[x_n^2 - 2x_n\mu + \mu^2] \tag{50}$$

$$= \frac{1}{N}\sum_{n=1}^{N} E[x_n^2] - E[2x_n\mu] + E[\mu^2] \tag{51}$$

$$= \frac{1}{N}\sum_{n=1}^{N} E[x_n^2] - 2E[x_n]E[x_n] + E[\mu^2] \tag{52}$$

$$= \frac{1}{N}\sum_{n=1}^{N} E[x_n^2] - E[x_n]^2 \tag{53}$$

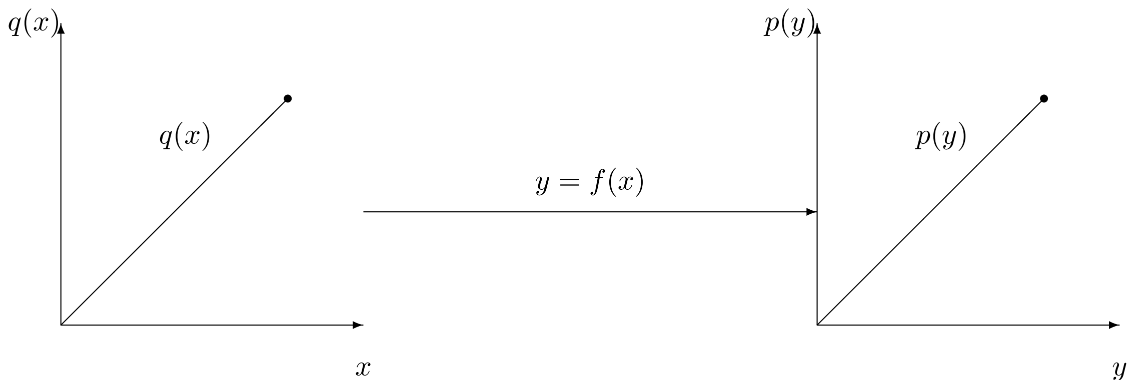
$$= \frac{1}{N}\sum_{n=1}^{N} \mu^2 + \sigma^2 - \mu^2 \tag{54}$$

$$= \sigma^2 \tag{55}$$

2.15 Exercise 2.18

Not attempted

2.16 Exercise 2.19



2.17 Exercise 2.20

Not attempted

2.18 Exercise 2.21

Showing $h(p^2) = 2h(p)$:

$$h(p) = h(p(x_1)) + h(p(x_2)) + \cdots + h(p(x_n)) \tag{56}$$
$$h(p^2) = h(x_1^2) + h(x_2^2) + \cdots + h(x_n^2) \tag{57}$$
$$\because h(x) = -\log_2 p(x), \tag{58}$$
$$h(p^2) = 2h(x_1) + 2h(x_2) + \cdots + 2h(x_n) \tag{59}$$
$$h(p^2) = 2h(p) \tag{60}$$

This can be applied to any exponent which inclues any choice of n integer or $\frac{n}{m}$ positive rational number.

$$h(p^x) = xh(p) \ \forall \ Q^+ \tag{61}$$
$$\therefore \tag{62}$$
$$h(p) \propto lnp \tag{63}$$

2.19 Exercise 2.22

Not attempted

2.20 Exercise 2.23

Not attempted

2.21 Exercise 2.24

Not attempted

2.22 Exercise 2.25

Not attempted

2.23 Exercise 2.26

Not attempted

2.24 Exercise 2.27

Not attempted

2.25 Exercise 2.28

Not attempted

2.26 Exercise 2.29

Not attempted

2.27 Exercise 2.30

Not attempted

2.28 Exercise 2.31

Not attempted

2.29 Exercise 2.32

Not attempted

2.30 Exercise 2.33

Not attempted

2.31 Exercise 2.34

Not attempted

2.32 Exercise 2.35

Not attempted

2.33 Exercise 2.36

Not attempted

2.34 Exercise 2.37

Not attempted

2.35 Exercise 2.38

Not attempted

2.36 Exercise 2.39

Not attempted

2.37 Exercise 2.40

Not attempted

2.38 Exercise 2.41

Not attempted

3 Standard Distributions

3.1 Exercise 3.1

Proving $\sum_{x=0}^{1} p(x|\mu) = 1$:

$$Bern(x|\mu) = \mu^x (1-\mu)^{1-x} \tag{64}$$
$$\sum_{x=0}^{1} p(x|\mu) = \mu^0 (1-\mu)^1 + \mu^1 (1-\mu)^0 \tag{65}$$
$$= 1 - \mu + \mu \tag{66}$$
$$= 1 \tag{67}$$

Proving $E[x] = \mu$:

$$E[x] = \sum_{x=0}^{1} p(x|\mu) * x \tag{68}$$
$$E[x] = 0 * \mu^0 (1-\mu)^1 + 1 * \mu^1 (1-\mu)^0 \tag{69}$$
$$= 0 * (1-\mu) + 1 * \mu \tag{70}$$
$$= \mu \tag{71}$$

Proving $var[x] = \mu(1-\mu)$:

$$Var[x] = \sum_i (x_i - \mu)^2 p(x_i) \tag{72}$$
$$= (0-\mu)^2 (1-\mu) + (1-\mu)^2 \mu \tag{73}$$
$$= \mu^2 (1-\mu) + (1-\mu)^2 \mu \tag{74}$$
$$= \mu(1-\mu)(\mu + (1-\mu)) \tag{75}$$
$$= \mu(1-\mu)(1) \tag{76}$$
$$= \mu(1-\mu) \tag{77}$$

Solving for entropy of the bernoulli distribution proves simple as there are two probabilities, $p_1 = \mu, p_2 = 1 - \mu$. Therefore the entropy must be:

$$H[x] = -\mu ln\mu - (1-\mu) ln(1-\mu) \tag{78}$$

3.2 Exercise 3.2

Not attempted

3.3 Exercise 3.3

Normalizing the binomial distribution $\binom{N}{m}\mu^m(1-\mu)^{N-m}$:

$$\text{First step is proving } \binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m} : \tag{79}$$

$$\binom{N}{m} + \binom{N}{m-1} = \frac{N!}{(N-m)!m!} + \frac{N!}{(N-m+1)!(m-1)!} \tag{80}$$

$$= \frac{N!}{(N-m)!m!} * \frac{N-m+1}{N-m+1} + \frac{N!}{(N-m+1)!(m-1)!} * \frac{m}{m} \tag{81}$$

$$= \frac{N!(N+1-m)}{(N+1-m)!m!} + \frac{N!m}{(N-m+1)!m!} \tag{82}$$

$$= \frac{(N+1)! - N!m + N!m}{(N+1-m)!m!} \tag{83}$$

$$= \frac{(N+1)!}{(N+1-m)!m!} \tag{84}$$

$$= \binom{N+1}{m} \tag{85}$$

With this information we would like to prove that $(1+x)^N = \sum_{m=0}^{N} \binom{N}{m} x^m$
N=2:

$$\binom{2}{0}x^0 + \binom{2}{1}x^1 + \binom{2}{2}x^2 = 1 + 2x + x^2 = (1+x)^2 \tag{86}$$

N=3:

$$\binom{3}{0}x^0 + \binom{3}{1}x^1 + \binom{3}{2}x^2\binom{3}{3}x^3 = 1 + 3x + 3x^2 + x^3 = (1+x)^3 \tag{87}$$

N=4

$$\binom{4}{0}x^0 + \binom{4}{1}x^1 + \binom{4}{2}x^2\binom{4}{3}x^3 + \binom{4}{4}x^4 = 1 + 4x + 6x^2 + 4x^3 + x^4 = (1+x)^4 \tag{88}$$

By induction: $\forall x \in \mathbb{R}$ and $N \in \mathbb{N}$

$$(1+x)^N = \sum_{m=0}^{N} \binom{N}{m} x^m \tag{89}$$

Normalizing the binomial distribution:

$$\sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} = 1 \tag{90}$$

$$= (1-\mu)^N \sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{-m} \tag{91}$$

$$= (1-\mu)^N \sum_{m=0}^{N} \binom{N}{m} (\frac{\mu}{1-\mu})^m \tag{92}$$

$$= (1-\mu)^N (1 + \frac{\mu}{1-\mu})^N \tag{93}$$

$$= ((1-\mu)(1 + \frac{\mu}{1-\mu}))^N \tag{94}$$

$$= (1 + \frac{\mu}{1-\mu} - \mu - \frac{\mu^2}{1-\mu}) \tag{95}$$

$$= ((1-\mu) + \frac{\mu}{1-\mu}(1-\mu))^N \tag{96}$$

$$= (1-\mu+\mu)^N \tag{97}$$

$$= 1 \tag{98}$$

3.4 Exercise 3.4

Not attempted

3.5 Exercise 3.5

Finding the mode of the multivariate gaussian. The mode of a distribution is the same as the maximum of the probability density function. For the multivariate gaussian, this is:

$$\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)} \tag{99}$$

$$\frac{d}{d\mathbf{x}}\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) \text{ ignoring normalization constant for brevity} \tag{100}$$

$$0 = -\frac{1}{2} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)} * \frac{d}{d\mathbf{x}}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) \tag{101}$$

$$0 = -\frac{1}{2} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)} * 2\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) \tag{102}$$

$$\mathbf{x} = \mu \tag{103}$$

3.6 Exercise 3.6

Not attempted

3.7 Exercise 3.7

Finding Kullback-Leibler divergence between $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu_{\mathbf{q}}, \mathbf{\Sigma_q})$ and $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu_{\mathbf{p}}, \mathbf{\Sigma_p})$

$$KL(p||q) = -\int p(x) ln \frac{q(x)}{p(x)} dx \tag{104}$$

$$KL(q(x)||p(x)) = -\int \mathcal{N}(\mathbf{x}|\mu_{\mathbf{q}}, \mathbf{\Sigma_q}) ln\frac{\mathcal{N}(\mathbf{x}|\mu_{\mathbf{p}}, \mathbf{\Sigma_p})}{\mathcal{N}(\mathbf{x}|\mu_{\mathbf{q}}, \mathbf{\Sigma_q})} \tag{105}$$

$$= -\int \mathcal{N}(\mathbf{x}|\mu_{\mathbf{q}}, \mathbf{\Sigma_q})(ln\mathcal{N}(\mathbf{x}|\mu_{\mathbf{p}}, \mathbf{\Sigma_p}) - ln\mathcal{N}(\mathbf{x}|\mu_{\mathbf{q}}, \mathbf{\Sigma_q})) \tag{106}$$

$$\text{definition: } ln\mathcal{N}(\mathbf{x}|\mu_{\mathbf{p}}, \mathbf{\Sigma_p}) = -\frac{D}{2}ln(2\pi) - \frac{1}{2}ln|\mathbf{\Sigma_p}| - \frac{1}{2}(\mathbf{x}-\mu_{\mathbf{p}})^T\mathbf{\Sigma_p}^{-1}(\mathbf{x}-\mu_{\mathbf{p}}) \tag{107}$$

$$= -\frac{1}{2}\int \mathcal{N}(\mathbf{x}|\mu_{\mathbf{q}}, \mathbf{\Sigma_q})(-ln|\mathbf{\Sigma_p}| + ln|\mathbf{\Sigma_q}| - (\mathbf{x}-\mu_{\mathbf{p}})^T\mathbf{\Sigma_p}^{-1}(\mathbf{x}-\mu_{\mathbf{p}}) + (\mathbf{x}-\mu_{\mathbf{q}})^T\mathbf{\Sigma_q}^{-1}(\mathbf{x}-\mu_{\mathbf{q}})) \tag{108}$$

$$\text{Splitting integral into three} \tag{109}$$

$$\text{First: } -\frac{1}{2}ln\frac{|\mathbf{\Sigma_q}|}{|\mathbf{\Sigma_p}|}\int \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma_q}|^{\frac{1}{2}}}e^{-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{q}})^T\mathbf{\Sigma_q}^{-1}(\mathbf{x}-\mu_{\mathbf{q}})} \tag{110}$$

$$\text{Second: } \frac{1}{2}\int \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma_q}|^{\frac{1}{2}}}e^{-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{q}})^T\mathbf{\Sigma_q}^{-1}(\mathbf{x}-\mu_{\mathbf{q}})}(\mathbf{x}-\mu_{\mathbf{p}})^T\mathbf{\Sigma_p}^{-1}(\mathbf{x}-\mu_{\mathbf{p}}) \tag{111}$$

$$\text{Third: } -\frac{1}{2}\int \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma_q}|^{\frac{1}{2}}}e^{-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{q}})^T\mathbf{\Sigma_q}^{-1}(\mathbf{x}-\mu_{\mathbf{q}})}(\mathbf{x}-\mu_{\mathbf{p}})^T\mathbf{\Sigma_p}^{-1}(\mathbf{x}-\mu_{\mathbf{p}}) \tag{112}$$

$$\text{Evaluating first integral: } -\frac{1}{2}ln\frac{|\mathbf{\Sigma_q}|}{|\mathbf{\Sigma_p}|} * 1 \tag{113}$$

$$\text{Evaluating second integral: } \frac{1}{2}(\mu_{\mathbf{p}}-\mu_{\mathbf{q}})^T\mathbf{\Sigma_p}^{-1}(\mu_{\mathbf{p}}-\mu_{\mathbf{q}}) \tag{114}$$

$$\text{Evaluating third integral: } -\frac{1}{2}(\mu_{\mathbf{p}}-\mu_{\mathbf{q}})^T\mathbf{\Sigma_p}^{-1}(\mu_{\mathbf{p}}-\mu_{\mathbf{q}}) + \frac{1}{2}Tr(\mathbf{\Sigma_p^{-1}\Sigma_q}) \tag{115}$$

I am not sure about this one it is quite hard

3.8 Exercise 3.8

Not attempted

3.9 Exercise 3.9

Not attempted

3.10 Exercise 3.10

Not attempted

3.11 Exercise 3.11

Not attempted

3.12 Exercise 3.12

Not attempted

3.13 Exercise 3.13

Not attempted

3.14 Exercise 3.14

Not attempted

3.15 Exercise 3.15

Not attempted

3.16 Exercise 3.16

Not attempted

3.17 Exercise 3.17

Not attempted

3.18 Exercise 3.18

Not attempted

3.19 Exercise 3.19

Not attempted

3.20 Exercise 3.20

Not attempted

3.21 Exercise 3.21

Not attempted

3.22 Exercise 3.22

Not attempted

3.23 Exercise 3.23

Not attempted

3.24 Exercise 3.24

Not attempted

3.25 Exercise 3.25

Not attempted

3.26 Exercise 3.26

Not attempted

3.27 Exercise 3.27

Not attempted

3.28 Exercise 3.28

Not attempted

3.29 Exercise 3.29

Not attempted

3.30 Exercise 3.30

Not attempted

3.31 Exercise 3.31

Not attempted

3.32 Exercise 3.32

Not attempted

3.33 Exercise 3.33

Not attempted

3.34 Exercise 3.34

Not attempted

3.35 Exercise 3.35

Not attempted

3.36 Exercise 3.36

Not attempted

3.37 Exercise 3.37

Deriving a maximimum likelihood estimator for a histogram-like density model. Space $\mathbf{x}$ is divided into fixed regions with density $p(\mathbf{x}) = h_i \; \forall \; i$ regions. Volume i is denoted Δ_i. N total observations such that n_i observations are in region i.

$$h_i = \frac{n_i}{N\Delta_i} \tag{116}$$

$$\sum_i h_i\Delta_i = 1 \tag{117}$$

$$L = \prod_{i=0}^{K} h_i^{n_i} \tag{118}$$

$$logL = \prod_{i=0}^{K} n_i logh_i \tag{119}$$

$$max \sum_{i=0}^{K} n_i logh_i - \lambda(\sum_i h_i\Delta_i - 1) \tag{120}$$

$$0 = \frac{n_i}{h_i} - \lambda\Delta_i \tag{121}$$

$$\frac{n_i}{h_i} = \lambda\Delta_i \tag{122}$$

$$h_i = \frac{n_i}{\lambda\Delta_i} \tag{123}$$

$$\text{Plugging into normalization:} \tag{124}$$

$$\sum_i \frac{n_i}{\lambda} = 1 \tag{125}$$

$$\lambda = N \tag{126}$$

$$h_i = \frac{n_i}{N\Delta_i} \tag{127}$$

3.38 Exercise 3.38

Not attempted

4 Single-layer Networks: Regression

4.1 Exercise 4.1

Showing the coefficients $\mathbf{w} = \{w_i\}$ that minimize the error function $E[\mathbf{w}] = \frac{1}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2$ are given by the solution to:

$$\sum_{j=0}^{M} A_{ij} w_j = T_i \tag{128}$$

where $A_{ij} = \sum_{n=1}^{N} x_n^{i+j}$, $T_i = \sum_{n=1}^{N} x_n^i t_n$ and $y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j$

$$\frac{1}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 = \frac{1}{2}\sum_{n=1}^{N}(\sum_{j=0}^{M} w_j x_n^j - t_n)^2 \tag{129}$$

$$\frac{d}{dw_i} = \sum_{n=1}^{N}\sum_{j=0}^{M}(w_j x_n^j - t_n)x_n^i \tag{130}$$

$$0 = \sum_{j=0}^{M} w_j \sum_{n=1}^{N} x_n^{i+j} - x_n^i t_n \tag{131}$$

$$\sum_{n=1}^{N} x_n^i t_n = \sum_{j=0}^{M} w_j \sum_{n=1}^{N} x_n^{i+j} \tag{132}$$

$$T_i = \sum_{j=0}^{M} A_{ij} w_j \tag{133}$$

4.2 Exercise 4.2

Not Attempted

4.3 Exercise 4.3

Showing that the *tanh* function is related to the logistic sigmoid function by $tanh(a) = 2\sigma(2a) - 1$

$$tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}} \tag{134}$$

$$\sigma(a) = \frac{1}{1 + e^{-a}} \tag{135}$$

$$\frac{e^a - e^{-a}}{e^a + e^{-a}} + 1 = 2\sigma(2a) \tag{136}$$

$$\frac{\frac{e^{2a}-1}{e^a}}{\frac{e^{2a}+1}{e^a}} + 1 = 2\sigma(2a) \tag{137}$$

$$\frac{e^{2a} - 1}{e^{2a} + 1} + 1 = 2\sigma(2a) \tag{138}$$

$$\frac{2e^{2a}}{e^{2a} + 1} = 2\sigma(2a) \tag{139}$$

$$\sigma(2a) = \frac{e^{2a}}{1 + e^{2a}} \tag{140}$$

Showing that $w_0 + \sum_{j=1}^{M} w_j \sigma(\frac{x-\mu_j}{s})$ is a linear combination of $u_0 + \sum_{j=1}^{M} u_j tanh(\frac{x-\mu_j}{2s})$
Defining $z_j = \frac{x-\mu_j}{s}$

$$u_0 + \sum_{j=1}^{M} u_j tanh(\frac{z_j}{2}) \tag{141}$$

$$= u_0 + -1 + 2\sum_{j=1}^{M} u_j \sigma(z_j) \tag{142}$$

$$y(x, \mathbf{u}) = u_0 - 1 - 2w_0 + 2y(x, \mathbf{w}) \tag{143}$$

4.4 Exercise 4.4

Not attempted

4.5 Exercise 4.5

The weighted sum of squares error function could attempt to weigh some data points more than others. This could tie to data-dependent noise variance by weighing the points that are outliers less than other points. The weighted part could also be a way to lower the weight of or completely negate the effect of duplicate data points. A statement could be set up where if a data point was duplicated, the weight will be set to 0.
Finding the solution $\mathbf{w}^*$ that minimizes

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n(t_n - \mathbf{w}^T\phi(\mathbf{x}_n))^2 \tag{144}$$

$$\frac{d}{d\mathbf{w}} = -\sum_{n=1}^{N} r_n \phi_n (t_n - \mathbf{w}^T \phi_n) \tag{145}$$

$$0 = \sum_{n=1}^{N} r_n \phi_n t_n - r_n \phi_n \mathbf{w}^T \phi_n \tag{146}$$

$$\sum_{n=1}^{N} \mathbf{w}^T \phi_n = \sum_{n=1}^{N} t_n \tag{147}$$

$$\mathbf{w} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t} \tag{148}$$

4.6 Exercise 4.6

Not attempted

4.7 Exercise 4.7

Finding Σ_{ml}:

$$p(t|W, \Sigma) = \mathcal{N}(t|y(x, W), \Sigma) \tag{149}$$

$$y(x, W) = W^T \phi(x) \tag{150}$$

likelihood function of multivariate target variable t is:

$$lnp(t|y(x, W), \Sigma) = -\frac{ND}{2} ln(2\pi) - \frac{N}{2} ln|\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (t_n - y(x, W))^T \Sigma^{-1} (t_n - y(x, W)) \tag{151}$$

$$= -\frac{ND}{2} ln(2\pi) - \frac{N}{2} ln|\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (t_n - W^T \phi(x))^T \Sigma^{-1} (t_n - W^T \phi(x)) \tag{152}$$

$$\frac{d}{d\Sigma} = -\frac{N}{2|\Sigma|} \frac{d}{d|\Sigma|} + \frac{1}{2} \sum_{n=1}^{N} (t_n - W^T \phi(x))^T \Sigma^{-1} I \Sigma^{-1} (t_n - W^T \phi(x)) \tag{153}$$

$$\frac{N}{2} \Sigma^{-1} = \Sigma^{-1} \frac{1}{2} \sum_{n=1}^{N} (t_n - W^T \phi(x))(t_n - W^T \phi(x))^T \Sigma^{-1} \tag{154}$$

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} (t_n - W_{ML}^T \phi(x))(t_n - W_{ML}^T \phi(x))^T \tag{155}$$

Important notes learned from this exercise: tr scalar = scalar and $X^T y X$ is a scalar for any dimension compatible matrices. Also, tr(ABC) = tr(BCA) = tr(CAB)

4.8 Exercise 4.8

Not attempted

4.9 Exercise 4.9

Finding $E[L(t, f(x))]$ for multiple target variables where:

$$E[L(t, f(x))] = \int \int ||f(x) - t||^2 p(x, t) dx dt \tag{156}$$

And showing that the function f(x) that minimizes E is given by $E_t[t|x]$

$$E[L(t, f(x))] = \int \int ||f(x) - t||^2 p(x, t) dx dt \tag{157}$$

$$= \int \int ||f(x) - E[t|x] + E[t|x] - t||^2 p(x, t) dx dt \tag{158}$$

$$= \int \int ||f(x) - E[t|x]||^2 + 2(f(x) - E[t|x])(E[t|x] - t) + ||E[t|x] - t||^2 p(x, t) dx dt \tag{159}$$

$$\text{Integrating over t} \tag{160}$$

$$\int ||f(x) - E[t|x]||^2 p(x) dx + \int var(t|x) p(x) dx \tag{161}$$

$$\text{Minimizing this gives} \tag{162}$$

$$f(x) = E[t|x] \tag{163}$$

4.10 Exercise 4.10

Not attempted

4.11 Exercise 4.11

Normalizing

$$p(x|\sigma^2, q) = \frac{q}{2(2\sigma^2)^{\frac{1}{q}} \Gamma(\frac{1}{q})} e^{-\frac{|x|^q}{2\sigma^2}} \tag{164}$$

where

$$\Gamma(x) = \int_{-\infty}^{\infty} u^{x-1} e^{-u} du \tag{165}$$

$$\Gamma(\frac{1}{q}) = \int_{-\infty}^{\infty} u^{\frac{1}{q}-1} e^{-u} du \tag{166}$$

$$\Gamma(\frac{1}{q}) = \int_{-\infty}^{\infty} u^{\frac{1}{q}-1} e^{-u} du \tag{167}$$

Not finished

4.12 Exercise 4.12

Not attempted