Brief Article

The Author

September 14, 2013

1 1.1

Straightforward. Calualte $\frac{\partial E(\mathbf{w})}{\partial w_j} = 0$ for $j = 0, \ldots, M$.

2 1.2

Straightforward. Like 1.1.

After reading the solution manual, I think it's much better to present A as a whole matrix.

3 1.3

$P(a) = 0.2 \times 3/10 + 0.2 \times 1/2 + 0.6 \times 3/10 = 0.34$. $P(g|o) = P(g)P(o|g)/P(o) = 0.6 * 0.3/0.36 = 0.5$.

4 1.4

By brute-force differentiation, we have

$$p'_y(y) = sp'_x(g)(g'(y))^2 + sp_x(g)g''(y) \tag{1}$$

Assuming $\widehat{x} = g(\widehat{y}), p'_x(\widehat{x}) = p'_y(\widehat{y}) = 0$, we have $p'_y(\widehat{y}) = sp'_x(\widehat{x})(g'(\widehat{y}))^2 + sp_x(\widehat{x})g''(\widehat{y}) = sp_x(\widehat{x})g''(\widehat{y}) = 0$, and this requires $g''(\widehat{y}) = 0$.

In general, this won't be satisfied. However, when $g(\cdot)$ is linear, we have $g''(\widehat{y}) = 0$.

In this question, the uniqueness about $\widehat{x}$ and $\widehat{y}$ is ignored, but this is irrelevant in most cases.

5 1.5

Straightforward.

6 1.6

Use the fact $p(x, y) = p(x)p(y)$ and change the order of integration (or integrate x and y separately)

7 1.7

Here, I just use the result $I = (2\pi\sigma^2)^{1/2}$. Then just use this result in the integration of (1.46), and we are done.

8 1.8

About (1.49): well... I have to admit that this problem is not as complicated as I thought... Let $y = x - \nu$ and notice it's the sum of the integration of a odd function, and the integration of a constant times the integration in 1.7.

About (1.50): I'm so poor at calculus...

9 1.9

Trivial... Perhaps we need that Σ^{-1} is positive-definite...

10 1.10

Trivial...

11 1.11

Trivial...

12 1.12

Trivial...

13 1.13

Trivial...

14 1.14

Trivial... Let $w_{ij}^S = (w_{ij} + w_{ji})/2$, and $w_{ij}^A = (w_{ij} - wji)/2$.

15 1.15

(1.134): given original $x_{i_1}, x_{i_2}, \ldots, x_{i_M}$ (D^M in total), we can always arrange $i_1, i_2, \ldots$ in decreasing order. So, (1.134) should suffice.

(1.135): for each $\widetilde{w}$, consider the corresponding arrangement of $i_1, i_2, \ldots, i_M$. If $i_1 = D$, then other i_m ($M-1$ terms) must be less or equal than D, so for this case, there should be $n(D, M-1)$ possible arrangements of $i_1, i_2, \ldots, i_M$. Similarly, considering the case $i_1 = m$, there're $n(m, M-1)$ terms. So we have (1.135).

(1.136): when $D = 1$, it's obviously true. Then it's trivial...

(1.137): obvious for $M = 2$. For $M > 2$, use (1.136) and see that its form is just (1.135), so we're done.

16 1.16

(1.138): trivial.

(1.139): trivial.

(1.140): for each case ($D \ggg M$ or $M \ggg D$), set M or D constant, and expand using Stirling's formula.

Hint: you can multiply the intermediate result by M^D (the thing you desire) and divide it by M^D again. Things can become clear then.

17 1.17

trivial.. just remember $\int u\mathrm{d}v = uv - \int v\mathrm{d}u$

18 1.18

(1.142): for left side, use (1.126). for right side, use (1.141) and use $2r\mathrm{d}r = \mathrm{d}r^2$.

(1.144): here, we conveniently assume for a sphere of radius r, $S_D(r) = S_D(1)r^{D-1}$, and V_D is calculated by summing the volumes of many small sphere shells ($S_D(r)\mathrm{d}r$).

19 1.19

(1.145): volume of cube is $(2a)^D$. Letting $a = 1$, we have (1.145).

ratio of (1.146): trivial.

ratio of that in the text: distance to sides is a, and distance from center to corner is $\sqrt{Da^2}$, and the ratio $\sqrt{D}$ follows.

20 1.20

skip for now...

(1.148): I think (45) in reference solution manual is wrong. (1.148) is somewhat obvious. $p(r)$ should be $p(\mathbf{x})$ integrated over a specific sphere of radius r, and this sphere has area $S_D r^{D-1}$, so we get (1.148).

stationary point: differentiate $p(r)$ with respect to r.

stationary point plus ϵ: a lot of approximation... see reference solution.

density at origin and that at $\widehat{r}$: trivial.

21 1.21

$a \leq (ab)^{1/2}$: trivial.

The inequality: $p(\text{mistake})$ is an integration over $\mathbf{x}$, and so is the right side. At every $\mathbf{x}$, $p(\text{mistake})$ is the smaller (a) of $p(\mathbf{x}, \mathcal{C}_1)$ and $p(\mathbf{x}, \mathcal{C}_2)$). Then we're done.

22 1.22

Trivial. Interpretation: classification rate.

23 1.23

Trivial.

24 1.24

The phrase "decision criterion" means a rule to follow when making decisions. Given $\mathbf{x}$, with true label unknown, if we reject it, the loss is λ. If we classify it as class j, the loss is L_{kj}, assuming that its label is k. However, we don't know its true label, and the distribution of this is given by $p(\mathcal{C}|\mathbf{x})$, so the expected loss if we classify it as class j is $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$. If the minimum of this value (over j) is greater than λ, we reject. Otherwise, we choose j.

If $L_{kj} = 1 - I_{kj}$, then $\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) = 1 - p(\mathcal{C}_j|\mathbf{x})$. Therefore, the criterion becomes: if the minimum of $1 - p(\mathcal{C}_j|\mathbf{x})$ over j is greater than λ, we reject. Otherwise, we choose j. Here, $p(\mathcal{C}_j|\mathbf{x})$ is largest. So the criterion can be reformulated as: if largest of $p(\mathcal{C}_j|\mathbf{x})$ is less than $1 - \lambda$, we reject. So we have $\theta = 1 - \lambda$.

25 1.25

Skip...

26 1.26

Trivial...

27 1.27

Skip...

28 1.28

I think the wording and notation in page 48 is somewhat confusing. In my understanding, $h(\cdot)$ is a univariate function taking the probability of event x. So, $h(x, y)$ should be written as $h(p(x, y)) = h(p(x)p(y)) = h(p(x)) + h(p(y))$.

Using this notation, we first see $h(p^2) = h(p)h(p)$. Then by induction we trivially have $h(p^n) = h(p^{n-1}) + h(p) = (n-1)h(p) + h(p) = nh(p)$. Then, regarding $p^{1/m}$ as a whole, we have $h(p^{n/m}) = nh(p^{1/m}) = (n/m)mh(p^{1/m}) = n/mh(p)$. By continuity, we have $h(p^x) = xh(p)$. Last, given two positive real numbers $p, q = p^x$, we have

$$\frac{h(q)}{\ln q} = \frac{xh(p)}{x \ln p} = \frac{h(p)}{\ln p}. \tag{2}$$

Thus, we have $h(p) \propto \ln p$.

Hint: sometimes, we should regard different things as the "unit" to be learned, like $p^{1/m}$ and p.

29 1.29

$$H(x) = -\sum_{i=1}^{M} p(x_i) \ln p(x_i) = \sum_{i=1}^{M} p(x_i) \ln(1/p(x_i)) \tag{3}$$

Since $\ln(\cdot)$ is concave, the sign in the (1.115) should be reversed.

$$H(x) \leq \ln \sum_{i=1}^{M} (p(x_i)/p(x_i)) = \ln M. \tag{4}$$

Hint: it's wrong to let $f(x) = -\ln(x)$. This will lead to $H(x) \geq -\ln \sum_{i=1}^{M} p(x_i)^2 \leq \ln M$, which is true and useless.

30 1.30

Trivial... Just need patience and carefulness.

31 1.31

Trivial... The solution in the manual seems redundnant...Use $\mathrm{KL}(p(x,y)\|p(x)p(y)) \geq 0$, and everything follows.

32 1.32

See the manual... I have no idea of Jacobian...

33 1.33

Trivial... But we have to assume that $x_1 \ln x_2 = 0$, whenever $x_1 = 0$ or $x_2 = 0$.

34 1.34

Skip...

35 1.36

Refer to the manual... It's so tricky...

36 1.37

$$H[\mathbf{x},\mathbf{y}] = -\iint p(\mathbf{x},\mathbf{y}) \ln p(\mathbf{x},\mathbf{y}) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \tag{5}$$

$$= -\iint p(\mathbf{x},\mathbf{y}) \ln(p(\mathbf{y}|\mathbf{x}) + p(\mathbf{x})) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \tag{6}$$

By separating the logarithm of joint probability, the two terms become $H(\mathbf{x})$ and $H(\mathbf{y}|\mathbf{x})$ respectively.

37 1.38

Trivial...