Errata

(Mathematical Introduction to Data Science by Sven A. Wegner)

June 14, 2025

- Page 8, Line −4:

$$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\sum_{i=1}^{n}x_iy_i-\overline{y}\sum_{i=1}^{n}x_i-\overline{x}\sum_{i=1}^{n}y_i+n\overline{x}\overline{y}$$

- Page 9, Line −3:

$$0=\sum_{i=1}^{n}(ax_i+b-y_i)=a\sum_{i=1}^{n}x_i+nb-\sum_{i=1}^{n}y_i=an\overline{x}+nb-n\overline{y},$$

- Page 13, Line 10: …constant random variable av.
- Page 13, Line −5: since $\overline{x^{(n)2}}=\operatorname{var}(x^{(n)})+\overline{x^{(n)}}^{2}$ as the sum …
- Page 15, Line −9: $\cdots+\frac{2}{n^2}\sum_{i<j}x_ix_j\operatorname{E}(\mathcal{E}_i)\operatorname{E}\mathcal{E}_j)$
- Page 16, Line 10: …, i.e., $f^*(x)=\langle a^*,x\rangle+b^*$ …
- Page 14, Line −2: If the latter is the case, then $\operatorname{sign}(r_{xy})=\operatorname{sign}(\langle u,v\rangle)=\cdots$
- Page 17, Line 13: …we calculate (with just for now $\phi(\tilde{a})=\langle\tilde{a},X^{\mathrm{T}}X\tilde{a}\rangle$):
- Page 23, Line 9: In the picture it must be the z-axis.
- Page 23, Line −3: …and $(w,b)=(w_1,\dots,w_d,b)$ for …
- Page 24, Line 16: $\operatorname{P}\big[Y_i\cancel{(f)}=y_i \text{ for all } i\big]=\cdots$
- Page 24, Line 18: $L\colon\big\{f\colon\mathbb{R}^d\to(0,1)\,|\,f \text{ logistic function}\big\}\to\mathbb{R}$
- Page 26, Line 13: $h'(t)=\operatorname{sig}(t)\cancel{+1}$
- Page 27, Line 18: …the data *overlaps* if for every $w\in\mathbb{R}^{d+1}\backslash\{0\}$ there exists …
- Page 27, Line 22: …, the rounded ~~logisitic~~ logistic regressor, …
- Page 27, Line −10: if $\operatorname{sig}(\langle w,\widehat{x}\rangle)\geqslant 1/2$
- Page 37, Line 1:

```
function k-NN Classifier(D, k, x)
    D′ ← D,  A ← ∅
    for j ← 1 to k do
        z* ← argmin_{z∈D′} ρ(x, π_1(z))
        A ← A ∪ {z*},  D′ ← D′ \ {z*}
    for y in Y do
        N(y) ← #{a ∈ A | π_2(a) = y}
    ℓ ← argmax_{y∈Y} N(y)
    return ℓ
```

Here, $\pi_2(x,y)=y$ denotes the projection onto the second entry of $(x,y)\in D$ ~~and y^* is the label of z^*~~.

- Page 39, Line 9: …*and* $k<n$.
- Page 39, Line 10: *The calculation of the k-neareast neighbors of x can be implemented such that at most $(n\cdot d\cdot k)$-many mulitplikations have to be carried out.*
- Page 39, Line 14: In the Euklidean metric, it requires $(d-1)$-many multiplications to compute one distance if we omit the root, which we can do as it does not change the argmin. This leads to

$$(d-1)\cdot(n+(n-1)+\cdots+(n-k+1))\leqslant\cancel{C}\,d\cdot k\cdot n$$

multiplikationen ~~with a suitable $C\in\mathbb{N}$~~.

- Page 40, Line 3: …, we choose k-nearest neighbors $x_1,\dots,x_k$ of x and denote their labels by $y_1,\dots,y_k$.

- Page 40, Line 9: $f\colon X \to Y,\ f(x) = \dfrac{\sum_{i=1}^{k} w(x_i, x)\cdot y_i}{\sum_{i=1}^{k} w(x_i, x)}$
- Page 41, Line 5: $\tilde{x}^{(i)} = \left(a + \dfrac{(x_1^{(i)} - \min_{j=1,\ldots,n} x_1^{(j)})(b-a)}{\max_{j=1,\ldots,n} x_1^{(j)} - \min_{j=1,\ldots,n} x_1^{(j)}}, \ldots\right)$
- Page 41, Line 7: $\tilde{x}^{(i)} = \left(\dfrac{x_1^{(i)} - \overline{x_1^{(\cdot)}}}{\sigma_1}, \ldots\right)$
- Page 42, Line 14: $\rho(x^{(1)}, x^{(4)}) = 3.681$
- Page 46, Line 1: We discuss some of these methods in Exercise 3.10.
- Page 45, Line 28: We thus see that texts no. 1 and text no. 2 are significantly more cosine similar than text no. 1 and text no. 3 or text no. 2 and text no. 3.
- Page 48, Line 26: The cosine distance, on the other hand, may appear here more natural, as the scalar product increases if the frequency of the fixed word increases in the second text.
- Page 52, Line 11: For finite subsets $A, B \subseteq X$ we define …
- Page 53, Line 20:

function Linkage-based Clustering (X, ρ, D, δ)
    $k \leftarrow \#D$
    for $i \leftarrow 1$ to k **do**
        $C_i \leftarrow \{x_i\}$
    while $\min_{i\neq j} \rho(C_i, C_j) \leqslant \delta$ and $k \geqslant 2$ **do**
        $m \leftarrow 0$
        $(i^*, j^*) \leftarrow \operatorname{argmin}_{i\neq j} \rho(C_i, C_j)$
        for $\ell \leftarrow 1$ to $k-1$ **do**
            if $\ell = \min(i^*, j^*)$ **then**
                $C_\ell \leftarrow C_{i^*} \cup C_{j^*}$
            if $\ell = \max(i^*, j^*)$ **then**
                $m \leftarrow 1$
                $C_\ell \leftarrow C_{\ell+m}$
            else
                $C_\ell \leftarrow C_{\ell+m}$
        $k \leftarrow k - 1$
    return $C_1, \ldots, C_k$

- Page 54, Line −6: $K\colon \mathcal{C}_k \to \mathbb{R}$
- Page 56, Line −16: *The following pseudocode approximates a minimizer of the k-means cost function.*
- Page 56, Pseudocode:

function k-means (D, k, X, ρ)
    $\mu_1, \ldots, \mu_k \leftarrow$ pairwise different points from X
    for $i \leftarrow 1$ to k **do**
        $C_i \leftarrow \{x \in D \mid i \in \operatorname{argmin}_{j=1,\ldots,k} \rho(x, \mu_j)\}$
    $U \leftarrow$ True
    while $U =$ True **do**
        $U \leftarrow$ False
        for $i \leftarrow 1$ to k **do**
            $\mu_i \leftarrow \mu(C_i)$
        for $i \leftarrow 1$ to k **do**
            $C_i' \leftarrow \{x \in D \mid i \in \operatorname{argmin}_{j=1,\ldots,k} \rho(x, \mu_j)\}$
            if $C_i' \neq C_i$
                $C_i \leftarrow C_i'$
                $U \leftarrow$True
    return $C_1, \ldots, C_k$

In the lines 4 and 9 of the pseudocode we pick as a single i in the case that the armin is not unique.

- Page 57, Line 7:

$$\mu(A) \in \underset{\mu\in A}{\operatorname{argmin}} \sum_{x\in A} \rho(x,\mu)^2, \quad \text{respectively} \quad \mu(A) \in \underset{\mu\in X}{\operatorname{argmin}} \sum_{x\in A} \rho(x,\mu).$$

- Page 57, Line -7: For $j \geqslant 1$ denote by $(C_1^{(j)}, \dots, C_k^{(j)})$ that clustering which the algorithm produces in the j-th round. For $j \geqslant 2$ we have

$$K(C_1^{(j)}, \dots, C_k^{(j)}) \;=\; \min_{\mu_1,\dots,\mu_k\in X} \sum_{i=1}^{k} \sum_{x\in C_i^{(j)}} \rho(x,\mu_i)^2$$

- Page 58, Line 3: ...line 11 of Algorithm 4.9 ...
- Page 58, Line 7: There, Picture 1b corresponds to the penultimate line in the estimate and Picture 2a to the line above that.
- Page 58, Line 10: ...as we have just moved point x_3 from cluster C_2 in Figure 1b to cluster C_1 in Figure 2a, ...
- Page 59, Line -4: We assume that we start with the initial values $\mu_1 = 2$ and ...
- Page 62, Line 11: $A = (a_{ij})_{i,j=1,\dots,n}$
- Page 62, Line 13: $L = (\ell_{ij})_{i,j=1,\dots,n}$
- Page 63, Line -2: ...In Example 5.7, $\lambda_2 \neq 0$ and there are no clusters (or, depending on how one prefers to see it, one single cluster), in Example 5.8 ...
- Page 67, Line 8: For the other direction let $\{v_1, \dots, v_n\}$ be a basis consisting of eigenvectors corresponding to the λ_i and let $U \subseteq \mathbb{R}^n$ be a subspace with $\dim U = n-k+1$. By construction $U \cap \operatorname{span}\{v_1,\dots,v_k\} \neq \{0\}$ and we can select $0 \neq x = \alpha_1 v_1 + \cdots + \alpha_k v_k \in U$. Then it follows

$$\frac{\langle x, Mx\rangle}{\langle x, x\rangle} = \frac{\sum\limits_{i=1}^{k} \lambda_i \alpha_i^2}{\sum\limits_{i=1}^{k} \alpha_i^2} \leqslant \frac{\lambda_k \sum\limits_{i=1}^{k} \alpha_i^2}{\sum\limits_{i=1}^{k} \alpha_i^2} = \lambda_k,$$

since the λ_i's are increasing. With this we get $\min_{0\neq x\in U} \frac{\langle x,Mx\rangle}{\langle x,x\rangle} \leqslant \lambda_k$ which then leads to

$$\max_{\substack{U\subseteq\mathbb{R}^n\\ \dim U = n-k+1}} \; \min_{\substack{x\in U\\ x\neq 0}} \frac{\langle x, Mx\rangle}{\langle x, x\rangle} \leqslant \lambda_k.$$

- Page 67, Line -4: ...and any corresponding eigenvector v_2 of norm 1:
- Page 68, Line 14: Let $G = (V,E)$ be a graph with $\deg(v) > 0$ for all $v \in V$.
- Page 70, Line 10:

$$\lambda_2(\mathcal{L}) = \min_{\substack{x\neq 0\\ \langle Dx,\mathbb{1}\rangle = 0}} \frac{\sum\limits_{\{i,j\}\in E} (x_i - x_j)^2}{\sum\limits_{i=1}^{n} x_i^2 d_i}.$$

- Page 70, Line 16:

$$\mathcal{L} D^{1/2}\mathbb{1} = D^{-1/2} L D^{-1/2} D^{1/2}\mathbb{1} = D^{-1/2} L \mathbb{1} \underset{\substack{\uparrow\\ \text{Proposition}\\ 5.6}}{=} D^{-1/2} 0 \mathbb{1} = 0 D^{1/2}\mathbb{1}.$$

- Page 70, Line 18:

$$\begin{aligned}\lambda_2(\mathcal{L}) &\overset{\substack{\text{Theorem}\\ 5.13}}{=} \min_{\substack{x\neq 0\\ \langle x, D^{1/2}\mathbb{1}\rangle = 0}} \frac{\langle x, \mathcal{L}x\rangle}{\langle x,x\rangle} \overset{(*)}{=} \min_{\substack{y\neq 0\\ \langle D^{1/2}y, D^{1/2}\mathbb{1}\rangle = 0}} \frac{\langle D^{1/2}y, \mathcal{L}D^{1/2}y\rangle}{\langle D^{1/2}y, D^{1/2}y\rangle}\\ &= \min_{\substack{y\neq 0\\ \langle y, D\mathbb{1}\rangle = 0}} \frac{\langle y, Ly\rangle}{\langle y, Dy\rangle} = \min_{\substack{y\neq 0\\ \langle Dy, \mathbb{1}\rangle = 0}} \frac{\sum_{\{i,j\}\in E}(y_i - y_j)^2}{\sum_{i=1}^{n} y_i^2 d_i}\end{aligned}$$

- Page 71, Line 14: *(ii)* $\min(\operatorname{vol} S_k, \operatorname{vol} S_k^{\mathrm{c}}) = \operatorname{vol} S_k^{\mathrm{c}}$ *and* $\operatorname{vol} S_k^{\mathrm{c}} - \operatorname{vol} S_{k+1}^{\mathrm{c}} = d_{k+1}$ *hold whenever* $r \leqslant k \leqslant n-1$.
- Page 71, Line 17:

$$\operatorname{vol} S_k^{\mathrm{c}} - \operatorname{vol} S_{k+1}^{\mathrm{c}} = \sum_{i=k+1}^{n} d_i - \sum_{i=k+2}^{n} d_i = d_{k+1}$$

- Page 72, Line 9:

$$\langle Dx, \mathbb{1} \rangle = \sum_{i=1}^{n} d_i x_i = \dots$$

- Page 72, Line -1 (and Page 73, Line 1):

$$\begin{aligned} \dots &= \operatorname{vol} S - \operatorname{vol} S \cdot \tfrac{\operatorname{vol} S}{\operatorname{vol} S + \operatorname{vol} S^{\mathrm{c}}} \\ &\geqslant \operatorname{vol} S - \operatorname{vol} S \cdot \tfrac{\operatorname{vol} S}{2 \operatorname{vol} S}, \end{aligned}$$

- Page 73, Line 9: …and our goal in the following will be to show $\lambda_2 \geqslant \alpha^2/2$, …
- Page 73, Line 12 (Equation (5.2)):

$$\cdots \text{ and } \langle Dx, \mathbb{1} \rangle = \sum_{i=1}^{n} d_i x_i = 0$$

- Page 73, Line 21:

$$\begin{bmatrix} x_1 - x_r \\ \vdots \\ x_{r-1} - x_r \\ 0 \\ x_{r+1} - x_r \\ \vdots \\ x_n - x_r \end{bmatrix} = \begin{bmatrix} x_1 - x_r \\ \vdots \\ x_{r-1} - x_r \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ x_r - x_{r+1} \\ \vdots \\ x_r - x_n \end{bmatrix} =: p - n.$$

- Page 74, Line -3 (until top of page 75):

$$\begin{aligned} \lambda_2 &= \frac{\sum_{\{i,j\} \in E} (x_i - x_j)^2}{\sum_{i=1}^{n} x_i^2 d_i} \\ &\underset{(5.3)}{\overset{\uparrow}{\geqslant}} \frac{\sum_{\{i,j\} \in E} \big((p_i - p_j)^2 + (n_i - n_j)^2\big)}{\sum_{i=1}^{n} (p_i^2 + n_i^2) d_i} \\ &\overset{(5.4)}{=} \frac{\sum_{\{i,j\} \in E} (p_i - p_j)^2 + \sum_{\{i,j\} \in E} (n_i - n_j)^2}{\sum_{i=1}^{n} p_i^2 d_i + \sum_{i=1}^{n} n_i^2 d_i} \\ &\underset{(5.5)}{\overset{\uparrow}{\geqslant}} \min\Big(\frac{\sum_{\{i,j\} \in E} (p_i - p_j)^2}{\sum_{i=1}^{n} p_i^2 d_i}, \frac{\sum_{\{i,j\} \in E} (n_i - n_j)^2}{\sum_{i=1}^{n} n_i^2 d_i} \Big) \\ &= \min\Big(\frac{\sum_{\{i,j\} \in E} (p_i - p_j)^2}{\sum_{i=1}^{n} p_i^2 d_i} \cdot \frac{\sum_{\{i,j\} \in E} (p_i + p_j)^2}{\sum_{\{i,j\} \in E} (p_i + p_j)^2}, \dots \Big) \\ &=: \min\Big(\frac{Z}{N}, \dots \Big). \end{aligned}$$

- Page 75, Line 8:

$$N = \sum_{i=1}^{n} p_i^2 d_i \cdot \sum_{\{i,j\} \in E} (p_i + p_j)^2 \underset{(5.5)}{\overset{\uparrow}{\geqslant}} \sum_{i=1}^{n} p_i^2 d_i \cdot \sum_{\{i,j\} \in E} 2(p_i^2 + p_j^2)$$

- Page 76, Line 3:

$$\cdots \underset{\substack{\text{telescopic} \\ \text{sum}}}{\overset{\uparrow}{=}} \Big(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{1}_E(i,j) \sum_{k=i}^{j-1} \big(p_k^2 - p_{k+1}^2\big) \Big)^2$$

- Page 79, Line 13: Finally, we want to note that Theorem 5.21 together with Remark 5.17(i) provides an upper bound for the eigenvalue $\lambda_2(\mathcal{L})$.
- Page 82, Line −5: *Let a dataset* $D = \{x_1, \dots, x_n\} \subseteq \mathbb{R}^d$ *be given, ...*
- Page 83, Line 3:
$$\sum_{j=1}^{k} \|Xv_j\|^2 \geqslant \sum_{j=1}^{k} \|Xw_j\|^2$$
- Page 83, Line −8:
$$\{v_1, \dots, v_k\} \in \operatorname*{argmax}_{\substack{\{\tilde{v}_1,\dots,\tilde{v}_k\} \subseteq \mathbb{R}^d \\ \text{orthonormal} \\ \text{system}}} \sum_{j=1}^{n} \|X\tilde{v}_j\|^2.$$
- Page 84, Line 10: Let $D = \{x_1, \dots, x_n\} \subseteq \mathbb{R}^d$ be given ...
- Page 84, Line 13: for $j = 1, \dots, k$, where ...
- Page 85, Line −1: $\langle w_{k+1}, v_i\rangle = \langle w_{k+1}, \underbrace{v_i - \pi_W(v_i)}_{\perp W}\rangle + \langle w_{k+1}, \underbrace{\pi_W(v_i)}_{\in U}\rangle = 0 + 0 = 0$
- Page 86, Line 11: Let $D = \{x_1, \dots, x_n\} \subseteq \mathbb{R}^d$ be given...
- Page 87, Line −9: Determine, by using Theorem 6.3 and a Python package for maximizations under nonlinear constraints, a 1-best fitting subspace for ...
- Page 86, Line 16: ...hold if $r = \dim \operatorname{span} D > 0$. If $r = 0$, then all σ_k vanish.
- Page 86, Line 19: For $k \geqslant r+1$ we apply the equation in the proof of Lemma 6.2 twice, on the one hand for V_k with ONB $\{v_1, \dots, v_k\}$, and on the other for V_{k-1} with ONB $\{v_1, \dots, v_{k-1}\}$. Then, after rearranging, we obtain
$$\sigma_k^2 = \|Xv_k\|^2 = \sum_{i=1}^{n} \operatorname{dist}(x_i, V_{k-1})^2 - \sum_{i=1}^{n} \operatorname{dist}(x_i, V_k)^2.$$
As $V_r \subseteq V_{k-1} \subseteq V_k$ holds, Lemma 6.4 implies that both terms on the right hand side are zero. Thus, $\sigma_{r+1}, \dots, \sigma_d = 0$.
- Page 90, Line 8: $\|u\|^2 = \|Av\|^2 = \langle Av, Av\rangle = \cdots$
- Page 94, Line 6: ..., it may be easier to start with this matrix and proceed analologuously (notice that one then gets the u_i's first and has to compute the v_i's by applying A^{T}).
- Page 94, Line −9: $V \in \mathbb{R}^{d\times d}$
- Page 94, Line −7: Then precisely r-many of the σ_i are nonzero and these are the singular values of A ...
- Seite 96, Line 10: ...and let $\{v_1, \dots, v_r\}$ be some arbitrary orthonormal system ...
- Page 96, Line 14: By construction, the $v_1, \dots, v_r$ that we defined in the theorem form an orthonormal system.
- Page 96, Line 20:
$$\|Av\|^2 = \langle v, A^{\mathrm{T}}Av\rangle = \langle v, B\operatorname{diag}(\cdots)B^{\mathrm{T}}v\rangle = \langle B^{\mathrm{T}}v, \operatorname{diag}(\cdots)B^{\mathrm{T}}v\rangle = \sum_{j=1}^{r} \lambda_j^{\cancel{2}}(B^{\mathrm{T}}v)_j^2$$
In the rest of the proof everywhere the squares at the eigenvalues λ_j have to be removed!
- Page 98, Line 4: ...v_i is therefore an eigenvector of $A^{\mathrm{T}}A$ corresponding to σ_i^2 for $i = 1, \dots, k$.
- Page 99, Line −3: ~~Multiplication with V^{T} from the left in~~
- Page 99, Line 12: $v = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sigma_{k+1} \\ w_{k+1} \end{bmatrix} \Bigg/ \left\| \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \sigma_{k+1} \\ w_{k+1} \end{bmatrix} \right\|$
- Page 99, Line 14: ..., we want to discuss how the singular value decomposition ~~indeed~~ generalizes the orthogonal diagonalization theorem.

- Page 99, Line 16: **Proposition 7.14.** Let $A \in \mathbb{R}^{n\times n}$ be a symmetric and positive semi-definite matrix, ...

- Page 99, Line -3: Proposition 7.14 thus yields exactly an orthogonl diagonalization $V^{\mathrm{T}}AV = \operatorname{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$; but only under the additional assumption of A being positive semidefinite.

  Corollary 5.22. Let $G = (V, E)$ be a graph with $\deg(i) > 0$ for all $i \in V$. Then for the second smallest eigenvalue λ_2 of the normalized Laplace matrix of G the estimate $\lambda_2 \leqslant 2$ holds. □

- Page 102, Line 17: *Proof.* Let $B \in \mathbb{R}^{n\times d}$ be arbitrary. Put

$$V_k := \operatorname{span}\{v_1, \ldots, v_k\} \quad \text{and} \quad W := \operatorname{span}\{b_1, \ldots, b_n\} \quad \cancel{\text{and} \quad \check{V} := \operatorname{span}\{\check{a}_1, \ldots, \check{a}_n\},}$$

  where $b_i \in \mathbb{R}^d$ is the vector that has as entries those of the i-th row of B. According to Theorem 7.12, V_k is a k-best fitting subspace for $\{a_1, \ldots, a_n\}$. ~~We claim that $\check{V} = V_k$. By Theorem 7.17(ii), we have $\check{a}_i \in V_k$ for $i = 1, \ldots, n$ and thus $\check{V} \subseteq V_k$. The equality then follows from $\dim \check{V} = \operatorname{rk} \check{A} = k = \dim V_k$.~~ Consequently ~~we obtain for each $i = 1, \ldots, n$ the estimates~~

$$\begin{aligned} \|A - B\|_{\mathrm{F}}^2 &= \sum_{i=1}^{n} \|a_i - b_i\|^2 \underset{\substack{\uparrow\\ b_i \in W}}{\geqslant} \sum_{i=1}^{n} \|a_i - \pi_W(a_i)\|^2 \underset{\substack{\uparrow\\ V_k\ k\text{-best}\\ \text{fitting for}\\ a_1,\ldots,a_n}}{\geqslant} \sum_{i=1}^{n} \|a_i - \pi_{V_k}(a_i)\|^2 \\ &\underset{\substack{\uparrow\\ \text{Thm}\\ 7.17\text{(ii)}}}{=} \sum_{i=1}^{n} \|a_i - \check{a}_i\|^2 = \|A - \check{A}\|_{\mathrm{F}}^2 \end{aligned}$$

  which concludes the proof. □

- Page 105, Zeile -1 and Page 106, Line 1:

$$A = \underbrace{\begin{bmatrix} 0.07 & 0.29 & 0.32 & 0.51 & 0.66 & 0.18 & -0.23 \\ 0.13 & -0.02 & -0.01 & -0.79 & 0.59 & -0.02 & -0.06 \\ 0.68 & -0.11 & -0.05 & -0.05 & -0.24 & 0.56 & -0.35 \\ 0.15 & 0.59 & 0.65 & -0.25 & -0.33 & -0.09 & 0.11 \\ 0.41 & -0.07 & -0.03 & 0.10 & -0.02 & -0.78 & -0.43 \\ 0.07 & 0.73 & -0.67 & 0.00 & -0.00 & 0.00 & 0.00 \\ 0.55 & -0.09 & -0.04 & 0.17 & 0.17 & -0.11 & 0.78 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 12.4 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 9.5 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.3 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} 0.56 & 0.09 & 0.56 & 0.09 & 0.59 \\ -0.12 & 0.69 & -0.12 & 0.69 & 0.02 \\ -0.40 & -0.09 & -0.40 & -0.09 & 0.80 \\ \cancel{-0.41} & \cancel{-0.09} & \cancel{0.40} & \cancel{0.09} & \cancel{-0.80} \\ 0.51 & 0.48 & -0.51 & -0.48 & -0.00 \\ 0.48 & -0.51 & -0.48 & 0.51 & -0.00 \end{bmatrix}}_{V^{\mathrm{T}}}$$

$$= \begin{bmatrix} 0.07 & 0.29 & 0.32 \\ 0.13 & -0.02 & -0.01 \\ 0.68 & -0.11 & -0.05 \\ 0.15 & 0.59 & 0.65 \\ 0.41 & -0.07 & -0.03 \\ 0.07 & 0.73 & -0.67 \\ 0.55 & -0.09 & -0.04 \end{bmatrix} \begin{bmatrix} 12.4 & & \\ & 9.5 & \\ & & 1.3 \end{bmatrix} \begin{bmatrix} 0.56 & 0.09 & 0.56 & 0.09 & 0.59 \\ -0.12 & 0.69 & -0.12 & 0.69 & 0.02 \\ -0.40 & -0.09 & -0.40 & -0.09 & 0.80 \end{bmatrix}.$$

- Page 106, Zeile 6:

$$\check{A} = \begin{bmatrix} 0.15 & 1.97 & 0.15 & 1.97 & 0.56 \\ 0.92 & 0.01 & 0.92 & 0.01 & 0.94 \\ 4.84 & 0.03 & 4.84 & 0.03 & 4.95 \\ 0.36 & 4.03 & 0.36 & 4.03 & 1.20 \\ 2.92 & -0.00 & 2.92 & -0.00 & 2.98 \\ -0.34 & 4.86 & -0.34 & 4.86 & 0.65 \\ 3.92 & 0.02 & 3.92 & 0.02 & 4.00 \end{bmatrix} = \begin{bmatrix} 0.07 & 0.29 \\ 0.13 & -0.02 \\ 0.68 & -0.11 \\ 0.15 & 0.59 \\ 0.41 & -0.07 \\ 0.07 & 0.73 \\ 0.55 & -0.09 \end{bmatrix} \begin{bmatrix} 12.4 & \\ & 9.5 \end{bmatrix} \begin{bmatrix} 0.56 & 0.09 & 0.56 & 0.09 & 0.59 \\ -0.12 & 0.69 & -0.12 & 0.69 & 0.02 \end{bmatrix}$$

- Page 107, Line 15:

$$= \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} 0.07 & 0.29 & \cdots & -0.23 \\ 0.13 & -0.02 & & -0.06 \\ 0.68 & -0.11 & & -0.35 \\ 0.15 & 0.59 & & 0.11 \\ 0.41 & -0.07 & & -0.43 \\ 0.07 & 0.73 & & 0.00 \\ 0.55 & -0.09 & \cdots & 0.78 \end{bmatrix} \begin{bmatrix} 12.4 & & & \\ & 9.5 & & \\ & & 1.3 & \\ & & & \ddots \end{bmatrix} \begin{bmatrix} 0.56 & 0.09 & 0.56 & 0.09 & 0.59 \\ -0.12 & 0.69 & -0.12 & 0.69 & 0.02 \\ \vdots & & & & \vdots \\ 0.48 & -0.51 & -0.48 & 0.51 & 0.00 \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

- Page 107, Line -2:

$$u_2 = \quad 0.29 \cdot \text{Abbie} - 0.02 \cdot \text{Bailey} + \cdots -0.09 \cdot \text{Gladys},$$

- Page 109, Line 10:

$$
\begin{array}{r}
\text{Abbie} \\ \text{Bailey} \\ \text{Catherine} \\ \text{Darlene} \\ \\ \text{Elena} \\ \text{Fatima} \\ \text{Gladys}
\end{array}
\check{A} = \begin{bmatrix} \boxed{0.07 \quad 0.29} \\ \boxed{0.13 \quad -0.02} \\ \boxed{0.68 \quad -0.11} \\ \boxed{0.15 \quad 0.59} \\ \boxed{0.41 \quad -0.07} \\ \boxed{0.07 \quad 0.73} \\ \boxed{0.55 \quad -0.09} \end{bmatrix} \begin{bmatrix} 12.4 & \\ & 9.5 \end{bmatrix}
\begin{array}{c}
\begin{array}{ccccc} \text{Alien} & \text{Casablanca} & \text{Star Wars} & \text{Titanic} & \text{The Matrix} \end{array} \\
\begin{bmatrix} \boxed{\begin{matrix} 0.56 \\ -0.12 \end{matrix}} & \boxed{\begin{matrix} 0.09 \\ 0.69 \end{matrix}} & \boxed{\begin{matrix} 0.56 \\ -0.12 \end{matrix}} & \boxed{\begin{matrix} 0.09 \\ 0.69 \end{matrix}} & \boxed{\begin{matrix} 0.59 \\ 0.02 \end{matrix}} \end{bmatrix}
\end{array}
$$

- Page 109, Line −3: …and $\check{\mathcal{V}} = \{v_1, v_2\}$.
- Page 119, Line −7 …, we assume that the distribution has mean zero and variance 1.
- Page 120, Line 1 in eq. (8.2): $V(\|X\|^2) = V(X_1^2) + \cdots + V(X_d^2)$.
- Page 121, Line 4: …, we get $\mathrm{E}(S_d) = 0$
- Page 121, Line 13: $= 2\sqrt{d}\,\mathrm{E}(\|X\| - \sqrt{d})$
- Page 121, Line −2: **Theorem 8.2.** Let $X, Y \sim \mathcal{N}(0, 1, \mathbb{R}^d)$ be independent. Then *(i)* $\forall\, d \in \mathbb{N}\colon |\mathrm{E}(\|X - Y\| - \sqrt{2d})| \leqslant 1/\sqrt{2d}$,
- Page 122, Line 1: *(ii)* $\forall\, d \in \mathbb{N}\colon\ \mathrm{V}(\|X - Y\|) \leqslant 4$.
- Page 122, Line 3 in eq. (8.2): $d \cdot \big(\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} x^4 \exp(-x^2/2)\,\mathrm{d}x - \mathrm{V}(X_1)^2\big)$
- Page 125, Line −7: *(ii)* $\forall\, d \in \mathbb{N}\colon\ \mathrm{V}(\|X - Y\|) \leqslant 4$.
- Page 125, Line −6: Firstly, calculate $\mathrm{V}((X_i - Y_i)^2) = 8$ …
- Page 125, Line −5: Then infer $\mathrm{V}((X_i - Y_i)^2) = 8d$ …
- Page 140, Line −7: *Proof.* We put $Y_1 + \cdots + Y_d$ and …
- Page 140, last expression in eq. (10.1): $= \mathrm{e}^{-ta} \prod_{i=1}^{d} \mathrm{E}(\mathrm{e}^{tY_i})$
- Page 143, Lines 2 − 6:

$$
\begin{aligned}
\mathrm{P}\big[\,|\|X\| - \sqrt{d}\,| > \varepsilon\big] &\leqslant \mathrm{P}\big[\,|\|X\| - \sqrt{d}| \cdot (\|X\| + \sqrt{d}) \geqslant \varepsilon \cdot \sqrt{d}\,\big] \\
&= \mathrm{P}\big[\,|X_1^2 + \cdots + X_d^2 - d| \geqslant \varepsilon\sqrt{d}\,\big] \\
&= \mathrm{P}\big[\,|(X_1^2 - 1) + \cdots + (X_d^2 - 1)| \geqslant \varepsilon\sqrt{d}\,\big] \\
&= \mathrm{P}\big[\,\big|\tfrac{X_1^2 - 1}{2} + \cdots + \tfrac{X_d^2 - 1}{2}\big| \geqslant \tfrac{\varepsilon\sqrt{d}}{2}\,\big] \\
&= \mathrm{P}\big[\,|Y_1 + \cdots + Y_d| \geqslant a\,\big],
\end{aligned}
$$

- Page 143, Line 10: 1 The Y_i are ~~pairwise~~ independent according to Fact 10.1(i).
- Page 152, Line −11:

$$
\mathrm{P}\Big[\,\Big|\Big\langle \frac{u_i}{\|u_i\|}, \frac{u_j}{\|u_j\|} \Big\rangle\Big| \leqslant \varepsilon\,\Big] \geqslant 1 - \frac{2/\varepsilon + 7}{\sqrt{d}}.
$$

- Page 155, Line 5: …(i) The dimension d of the ambient space …
- Page 184, Line −6: $y_i\langle w^{(j+1)}, \widehat{x}_i\rangle = y_i\langle w^{(j)} + y_i\widehat{x}_i, \widehat{x}_i\rangle = \underbrace{y_i\langle w^{(j)}, \widehat{x}_i\rangle}_{\substack{\text{erroneously} \\ \text{less than zero}}} + \underbrace{y_i^2\,\langle \widehat{x}_i, \widehat{x}_i\rangle}_{\substack{\text{addition of a} \\ \text{positive cor-} \\ \text{rection term}}}$.

- Page 185, Line 18: Since D is linearly separable, there exists some $w \in \mathbb{R}^{d+1} \backslash \{0\}$, such that all points will be classified correctly. ~~This vector w cannot be the zero vector, because then $h \equiv 0$ and all data points would be classified incorrectly.~~
- Page 185, Line -4: $\langle w^*, w^{(j+1)} \rangle \geqslant \langle w^*, w^{(0)} \rangle + (j+1) \cdot \gamma = (j+1) \cdot \gamma$
- Page 186, Line 3: $\|w^{(j+1)}\|^2 \leqslant \|w^{(0)}\|^2 + (j+1) \cdot R^2 = (j+1) \cdot R^2$.
- Page 186, Line 3: $(j+1) \cdot \gamma \langle w^*, w^{(j+1)} \rangle \leqslant \|w^*\|^{\cancel{2}} \|w^{(j+1)}\|^{\cancel{2}} = \|w^{(j+1)}\|^{\cancel{2}} + \sqrt{j+1} \cdot R$
- Page 229, Line 13: Alternatively, with sigmoid activation, ...
- Page 229, Line -4: $N = \begin{bmatrix} a_{11} & \cdots & a_{1\ell} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{m\ell} \end{bmatrix} \begin{bmatrix} n_1 \\ \vdots \\ n_\ell \end{bmatrix}$
- Page 229, Line -3: ... and a matrix $(a_{ij})_{i,j} \in \mathbb{R}^{m \times \ell}$ ...
- Page 237, Line 5: $\forall\, w \in \mathbb{R}^{\cancel{n}},\, b \in \mathbb{R}\colon \int_\Omega \sigma(wx + b) + \ell \,\mathrm{d}\mu(x) = 0$
- Page 237, Line -7: $\forall\, w \in \mathbb{R}^{\cancel{n}},\, b \in \mathbb{R}\colon \int_\Omega \sigma(wx + b) \,\mathrm{d}\mu(x) = 0$
- Page 242, Zeile -4: Since Ω is compact there exists $c_0 \geqslant 0$ such that $a_1 f_1 + \cdots + a_i f_i + c_0 \geqslant 0$ on Ω for all $i = 1, \ldots, \ell$. We construct ...
- Page 282, Line 6: (ii) For $A, B \in \Sigma$ with $\mathrm{P}(B) \neq 0$, $\mathrm{P}(A|B) := \dfrac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$ ...
- Page 283, Line 19:

$$\rho(x) = \tfrac{1}{(2\pi\sigma^2)^{d/2}}\, \mathrm{e}^{-\frac{\|x-\mu\|^2}{2\sigma^2}} \quad \text{respectively} \quad \rho(x) = \tfrac{1}{\lambda^d(B)} \cdot \mathbb{1}_B(x),$$

- Page 286, Line -5: For $A = A_1 \times \cdots \times A_d \subseteq \mathbb{R}^d$ with $A_i \in \mathcal{B}^{\cancel{d}}$ we calculate
- Page 288, Line 2:

$$\begin{aligned}
(\rho_1 * \rho_2)(s) &= \frac{1}{2\pi\sqrt{ab}} \int_\mathbb{R} \exp\big(-\frac{(s-t)^2}{2a}\big) \exp\big(-\frac{t^2}{2b}\big) \,\mathrm{d}t \\
&= \frac{1}{2\pi\sqrt{ab}} \int_\mathbb{R} \exp\big(-\frac{b(s^2 - 2st + t^2) + at^2}{2ab}\big) \,\mathrm{d}t \\
&= \frac{1}{2\pi\sqrt{ab}} \int_\mathbb{R} \exp\big(-\frac{t^2(b+a) - 2stb + bs^2}{2ab}\big) \,\mathrm{d}t \\
&= \frac{1}{2\pi\sqrt{ab}} \int_\mathbb{R} \exp\big(-\frac{t^2(b+a)/c - 2stb/c + bs^2/c}{2ab/c}\big) \,\mathrm{d}t \\
&= \frac{1}{\sqrt{2\pi c}\sqrt{2\pi(ab/c)}} \int_\mathbb{R} \exp\big(-\frac{(t-(bs)/c)^2 - (sb/c)^2 + s^2(b/c)}{2ab/c}\big) \,\mathrm{d}t \\
&= \frac{1}{\sqrt{2\pi c}} \exp\big(+\frac{(sb/c)^2 - s^2(b/c)}{2ab/c}\big) \frac{1}{\sqrt{2\pi(ab/c)}} \int_\mathbb{R} \exp\big(-\frac{(t-(bs)/c)^2}{2ab/c}\big) \,\mathrm{d}t \\
&= \frac{1}{\sqrt{2\pi c}} \exp\big(+\frac{(sb/c)^2 c^2 - s^2(b/c)c^2}{2abc}\big) \\
&= \frac{1}{\sqrt{2\pi c}} \exp\big(+\frac{s^2(b^2 - bc)}{2abc}\big) \\
&= \frac{1}{\sqrt{2\pi c}} \exp\big(-\frac{s^2}{2c}\big),
\end{aligned}$$