

Errata

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

July 6, 2025

- Page 8, Line -4:

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

- Page 9, Line -3:

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

- Page 13, Line 10: ... constant random variable **av**.
- Page 13, Line -5: since $\overline{x^{(n)^2}} = \text{var}(x^{(n)}) + \overline{x^{(n)}}^2$ as the sum ...
- Page 15, Line -9: $\dots + \frac{2}{n^2} \sum_{i < j} x_i x_j \mathbb{E}(\mathcal{E}_i) \mathbb{E} \mathcal{E}_j)$
- Page 16, Line 10: ..., i.e., $f^*(x) = \langle a^*, x \rangle + b^*$...
- Page 14, Line -2: If the latter is the case, then $\text{sign}(r_{xy}) = \text{sign}(\langle u, v \rangle) = \dots$
- Page 17, Line 13: ... we calculate (with just for now $\phi(\tilde{a}) = \langle \tilde{a}, X^T X \tilde{a} \rangle$):
- 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: $P[Y_i(\text{f}) = y_i \text{ for all } i] = \dots$
- Page 24, Line 18: $L: \{f: \mathbb{R}^d \rightarrow (0, 1) \mid f \text{ logistic function}\} \rightarrow \mathbb{R}$
- Page 26, Line 13: $h'(t) = \text{sig}(t) + 1$
- Page 27, Line 18: ... the data *overlaps* if for every $w \in \mathbb{R}^{d+1} \setminus \{0\}$ there exists ...
- Page 27, Line 22: ..., the rounded **logistie logistic** regressor, ...
- Page 27, Line -10: if $\text{sig}(\langle w, \hat{x} \rangle) \geq 1/2$
- Page 37, Line 1:

```
1: function K-NN CLASSIFIER( $D, k, x$ )
2:    $D' \leftarrow D, A \leftarrow \emptyset$ 
3:   for  $j \leftarrow 1$  to  $k$  do
4:      $z^* \leftarrow \text{argmin}_{z \in D'} \rho(x, \pi_1(z))$ 
5:      $A \leftarrow A \cup \{z^*\}, D' \leftarrow D' \setminus \{z^*\}$ 
6:   for  $y$  in  $Y$  do
7:      $N(y) \leftarrow \#\{a \in A \mid \pi_2(a) = y\}$ 
8:    $\ell \leftarrow \text{argmax}_{y \in Y} N(y)$ 
9:   return  $\ell$ 
```

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 -nearest neighbors of x can be implemented such that at most $(n \cdot d \cdot k)$ -many multiplications 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

$$(d - 1) \cdot (n + (n - 1) + \dots + (n - k + 1)) \leq \text{C} \cdot d \cdot k \cdot n$$

multiplicationen **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: $f: X \rightarrow Y, f(x) = \frac{\sum_{i=1}^k w(x_i, x) \cdot y_i}{\sum_{i=1}^k w(x_i, x)}$
- Page 41, Line 5: $\tilde{x}^{(i)} = \left(a + \frac{(x_1^{(i)} - \min_{j=1, \dots, n} x_1^{(j)})(b-a)}{\max_{j=1, \dots, n} x_1^{(j)} - \min_{j=1, \dots, n} x_1^{(j)}}, \dots \right)$
- Page 41, Line 7: $\tilde{x}^{(i)} = \left(\frac{x_1^{(i)} - x_1^{(-)}}{\sigma_1}, \dots \right)$
- 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:


```

1: function LINKAGE-BASED CLUSTERING ( $X, \rho, D, \delta$ )
2:    $k \leftarrow \#D$ 
3:   for  $i \leftarrow 1$  to  $k$  do
4:      $C_i \leftarrow \{x_i\}$ 
5:   while  $\min_{i \neq j} \rho(C_i, C_j) \leq \delta$  and  $k \geq 2$  do
6:      $m \leftarrow 0$ 
7:      $(i^*, j^*) \leftarrow \operatorname{argmin}_{i \neq j} \rho(C_i, C_j)$ 
8:     for  $\ell \leftarrow 1$  to  $k-1$  do
9:       if  $\ell = \min(i^*, j^*)$  then
10:         $C_\ell \leftarrow C_{i^*} \cup C_{j^*}$ 
11:       if  $\ell = \max(i^*, j^*)$  then
12:         $m \leftarrow 1$ 
13:         $C_\ell \leftarrow C_{\ell+m}$ 
14:       else
15:         $C_\ell \leftarrow C_{\ell+m}$ 
16:      $k \leftarrow k-1$ 
17:   return  $C_1, \dots, C_k$ 

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


```

1: function K-MEANS ( $D, k, X, \rho$ )
2:    $\mu_1, \dots, \mu_k \leftarrow$  pairwise different points from  $X$ 
3:   for  $i \leftarrow 1$  to  $k$  do
4:      $C_i \leftarrow \{x \in D \mid i \in \operatorname{argmin}_{j=1, \dots, k} \rho(x, \mu_j)\}$ 
5:    $U \leftarrow \text{True}$ 
6:   while  $U = \text{True}$  do
7:      $U \leftarrow \text{False}$ 
8:     for  $i \leftarrow 1$  to  $k$  do
9:        $\mu_i \leftarrow \mu(C_i)$ 
10:    for  $i \leftarrow 1$  to  $k$  do
11:       $C'_i \leftarrow \{x \in D \mid i \in \operatorname{argmin}_{j=1, \dots, k} \rho(x, \mu_j)\}$ 
12:      if  $C'_i \neq C_i$ 
13:         $C_i \leftarrow C'_i$ 
14:       $U \leftarrow \text{True}$ 
15:   return  $C_1, \dots, C_k$ 

```

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:

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

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

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

- 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 $0 \neq x = \alpha_1 v_1 + \dots + \alpha_k v_k \in U$. Then it follows

$$\frac{\langle x, Mx \rangle}{\langle x, x \rangle} = \frac{\sum_{i=1}^k \lambda_i \alpha_i^2}{\sum_{i=1}^k \alpha_i^2} \leq \frac{\lambda_k \sum_{i=1}^k \alpha_i^2}{\sum_{i=1}^k \alpha_i^2} = \lambda_k,$$

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

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

- 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:

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

- Page 70, Line 16:

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

- Page 70, Line 18:

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

- Page 71, Line 14: (ii) $\min(\text{vol } S_k, \text{vol } S_k^c) = \text{vol } S_k^c$ and $\text{vol } S_k^c - \text{vol } S_{k+1}^c = d_{k+1}$ hold whenever $r \leq k \leq \textcolor{red}{n} - 1$.

- Page 71, Line 17:

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

- Page 72, Line 9:

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

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

$$\begin{aligned} \dots &= \text{vol } S - \text{vol } \textcolor{red}{S} \cdot \frac{\text{vol } S}{\text{vol } S + \text{vol } S^c} \\ &\geq \text{vol } S - \text{vol } \textcolor{red}{S} \cdot \frac{\text{vol } S}{2 \text{vol } S}, \end{aligned}$$

- Page 73, Line 9: ... and our goal in the following will be to show $\lambda_2 \geq \alpha^2/\textcolor{red}{2}$, ...
- Page 73, Line 12 (Equation (5.2)):

$$\dots \text{ and } \langle \textcolor{red}{Dx}, \mathbf{1} \rangle = \sum_{i=1}^n d_i x_i = 0$$

- Page 73, Line 21:

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

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

$$\begin{aligned} \lambda_2 &= \frac{\sum_{\{i,j\} \in E} (x_i - x_j)^2}{\sum_{i=1}^n x_i^2 d_i} \\ &\geq \frac{\sum_{\{i,j\} \in E} ((p_i - p_j)^2 + (n_i - n_j)^2)}{\sum_{i=1}^n (p_i^2 + n_i^2) d_i} \\ &\stackrel{\substack{\uparrow \\ (5.3) \\ (5.4)}}{=} \frac{\sum_{\{i,j\} \in E} (p_i - p_j)^2 + \sum_{\{i,j\} \in E} (n_i - n_j)^2}{\sum_{i=1}^n p_i^2 d_i + \sum_{i=\textcolor{red}{1}}^n n_i^2 d_i} \\ &\stackrel{\substack{\uparrow \\ (5.5)}}{\geq} \min \left(\frac{\sum_{\{i,j\} \in E} (p_i - p_j)^2}{\sum_{i=1}^n p_i^2 d_i}, \frac{\sum_{\{i,j\} \in E} (n_i - n_j)^2}{\sum_{i=\textcolor{red}{1}}^n n_i^2 d_i} \right) \\ &= \min \left(\frac{\sum_{\{i,j\} \in E} (p_i - p_j)^2}{\sum_{i=1}^n p_i^2 d_i} \cdot \frac{\sum_{\{i,j\} \in E} (p_i + p_j)^2}{\sum_{\{i,j\} \in E} (p_i + p_j)^2}, \dots \right) \\ &=: \min \left(\frac{Z}{N}, \dots \right). \end{aligned}$$

- Page 75, Line 8:

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

- Page 76, Line 3:

$$\dots = \underset{\substack{\uparrow \\ \text{telescopic} \\ \text{sum}}}{\left(\sum_{i=1}^{n-1} \sum_{j=i+1}^n \mathbf{1}_E(i, j) \sum_{k=i}^{j-1} (p_k^2 - p_{k+1}^2) \right)^2}$$

- 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:

$$\sum_{j=1}^k \|Xv_j\|^2 \geq \sum_{j=1}^k \|Xw_j\|^2$$

- Page 83, Line -8:

$$\{v_1, \dots, v_k\} \in \underset{\substack{\{\tilde{v}_1, \dots, \tilde{v}_k\} \subseteq \mathbb{R}^d \\ \text{orthonormal} \\ \text{system}}}{\text{argmax}} \sum_{j=1}^n \|X\tilde{v}_j\|^2.$$

- 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 $r = \dim \text{span } D > 0$. If $r = 0$, then all σ_k vanish.
- Page 86, Line 19: For $k \geq r + 1$ 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

$$\sigma_k^2 = \|Xv_k\|^2 = \sum_{i=1}^n \text{dist}(x_i, V_{k-1})^2 - \sum_{i=1}^n \text{dist}(x_i, V_k)^2.$$

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: $\|u\|^2 = \|Av\|^2 = \langle Av, Av \rangle = \dots$
- Page 94, Line 6: ..., it may be easier to start with this matrix and proceed analogously (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 ...
- Page 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:

$$\|Av\|^2 = \langle v, A^T Av \rangle = \langle v, B \text{diag}(\dots) B^T v \rangle = \langle B^T v, \text{diag}(\dots) B^T v \rangle = \sum_{j=1}^r \lambda_j^2 (B^T v)_j^2$$

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 $A^T A$ corresponding to σ_i^2 for $i = 1, \dots, k$.
- Page 99, Line -3: ~~Multiplication with V^T from the left in~~

$$\text{■ Page 99, Line 12: } v = \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ \sigma_{k+1} \\ w_{k+1} \end{array} \right] \Bigg/ \left\| \left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ \sigma_{k+1} \\ w_{k+1} \end{array} \right] \right\|$$

- 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 orthogonal diagonalization $V^T A V = \text{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0)$; 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 $\lambda_2 \leq 2$ holds. \square

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

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

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 $\{a_1, \dots, a_n\}$. We claim that $\check{V} = V_k$. By Theorem 7.17(ii), we have $\check{a}_i \in V_k$ for $i = 1, \dots, n$ and thus $\check{V} \subseteq V_k$. The equality then follows from $\dim \check{V} = \text{rk } \check{A} = k = \dim V_k$. Consequently we obtain for each $i = 1, \dots, n$ the estimates

$$\begin{aligned} \|A - B\|_F^2 &= \sum_{i=1}^n \|a_i - b_i\|^2 \geq \sum_{i=1}^n \|a_i - \pi_W(a_i)\|^2 \geq \sum_{i=1}^n \|a_i - \pi_{V_k}(a_i)\|^2 \\ &\quad \uparrow \text{Thm 7.17(ii)} \\ &= \sum_{i=1}^n \|a_i - \check{a}_i\|^2 = \|\check{A} - A\|_F^2 \end{aligned}$$

V_k k -best fitting for a_1, \dots, a_n

which concludes the proof. \square

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

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

- Page 106, Line 6:

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

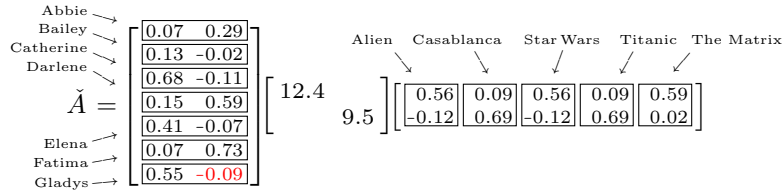
- Page 107, Line 15:

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

- Page 107, Line -2:

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

- Page 109, Line 10:



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

- Page 143, Lines 2 - 6:

$$\begin{aligned}
 P[|\|X\| - \sqrt{d}| > \varepsilon] &\leq P[|\|X\| - \sqrt{d}| \cdot (\|X\| + \sqrt{d}) \geq \varepsilon \cdot \sqrt{d}] \\
 &= P[|X_1^2 + \dots + X_d^2 - d| \geq \varepsilon \sqrt{d}] \\
 &= P[|(X_1^2 - 1) + \dots + (X_d^2 - 1)| \geq \varepsilon \sqrt{d}] \\
 &= P\left[\left|\frac{X_1^2 - 1}{2} + \dots + \frac{X_d^2 - 1}{2}\right| \geq \frac{\varepsilon \sqrt{d}}{2}\right] \\
 &= P[|Y_1 + \dots + Y_d| \geq a],
 \end{aligned}$$

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

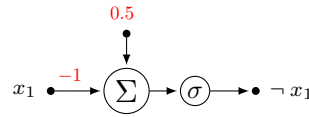
$$P\left[\left|\left\langle \frac{u_i}{\|u_i\|}, \frac{u_j}{\|u_j\|} \right\rangle\right| \leq \varepsilon\right] \geq 1 - \frac{2/\varepsilon + 7}{\sqrt{d}}.$$

- Page 155, Line 5: ... (i) The dimension d of the **ambient** space ...
- Page 155, Line -3ff:

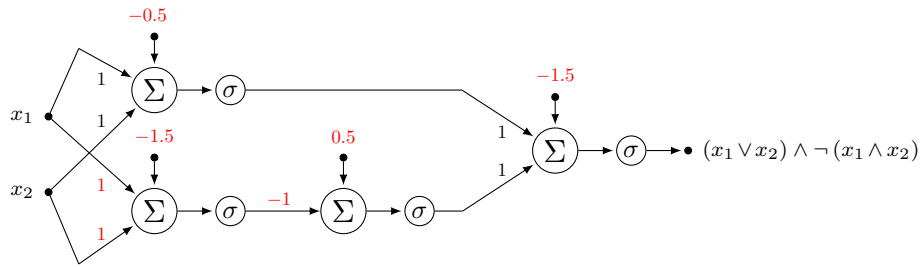
$$P[\dots] = 1 - P[\exists i \neq j: \|T_U x_i - T_U x_j\| \notin ((1-\varepsilon)\|x_i - x_j\|, (1+\varepsilon)\|x_i - x_j\|)]$$

$$\begin{aligned}
&\geq 1 - \sum_{i \neq j} (1 - P_{ij}) \geq 1 - \sum_{i \neq j} 2e^{-k\varepsilon^2/16} \geq 1 - \frac{n^2}{2} \cdot 2e^{-k\varepsilon^2/16} \\
&\geq 1 - n^2 e^{-(\frac{48}{\varepsilon^2} \log n) \varepsilon^2/16} = 1 - n^2 e^{\log(n^{-3})} \\
&= 1 - \frac{1}{n}
\end{aligned}$$

- Page 169, Line -1: $\sqrt{(\Delta(d))^2 + 2d} - \sqrt{2d} \rightarrow \infty$
- Page 170, Line -1: $\lim_{d \rightarrow \infty} \sqrt{(\Delta(d))^2 + 2d} - \sqrt{2d} = \infty$
- Page 171, Line 4: $\sqrt{(\Delta(d))^2 + 2d} - \sqrt{2d} \geq \sqrt{R\sqrt{d} + 2d} - \sqrt{2d} \xrightarrow{d \rightarrow \infty} \frac{R}{2\sqrt{2}}$
- Page 171, Line 5: Multiplication with $\sqrt{(\Delta(d))^2 + 2d} + \sqrt{2d} \dots$
- Page 173, Line 6: $\dots \geq R \cdot (\sqrt{(\Delta(d))^2 + 2d} - \sqrt{2d}) \geq \dots$
- Page 184, Line -6: $y_i \langle w^{(j+1)}, \hat{x}_i \rangle = y_i \langle w^{(j)} + y_i \hat{x}_i, \hat{x}_i \rangle = \underbrace{y_i \langle w^{(j)}, \hat{x}_i \rangle}_{\text{erroneously less than zero}} + \underbrace{y_i^2 \langle \hat{x}_i, \hat{x}_i \rangle}_{\text{addition of a positive correction term}}.$
- Page 185, Line 18: Since D is linearly separable, there exists some $w \in \mathbb{R}^{d+1} \setminus \{0\}$, 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: $\langle w^*, w^{(j+1)} \rangle \geq \langle w^*, w^{(0)} \rangle + (j+1) \cdot \gamma = (j+1) \cdot \gamma$
- Page 186, Line 3: $\|w^{(j+1)}\|^2 \leq \|w^{(0)}\|^2 + (j+1) \cdot R^2 = (j+1) \cdot R^2.$
- Page 186, Line 3: $(j+1) \cdot \gamma \langle w^*, w^{(j+1)} \rangle \leq \|w^*\|^2 \|w^{(j+1)}\|^2 = \|w^{(j+1)}\|^2 + \sqrt{j+1} \cdot R$
- Page 186, Line 1: under the underbrace: ≤ 0 , as \dots
- Page 187, Line -8: Assume that the perceptron algorithm outputs the **generalized** weight vector \dots
- Page 190, Line 2: and ~~let~~ $b \in \mathbb{R}$
- Page 227, Line 11:



- Page 227, Line -3:



- Page 229, Line 13: Alternatively, with **sigmoid** activation, \dots
- Page 229, Line -4: $N = \begin{bmatrix} a_{11} & \dots & a_{1\ell} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{m\ell} \end{bmatrix} \begin{bmatrix} n_1 \\ \vdots \\ n_\ell \end{bmatrix}$
- Page 229, Line -3: \dots and a matrix $(a_{ij})_{i,j} \in \mathbb{R}^{m \times \ell} \dots$
- Page 237, Line 5: $\forall w \in \mathbb{R}^{\text{??}}, b \in \mathbb{R}: \int_{\Omega} \sigma(wx + b) + \ell d\mu(x) = 0$
- Page 237, Line -7: $\forall w \in \mathbb{R}^{\text{??}}, b \in \mathbb{R}: \int_{\Omega} \sigma(wx + b) d\mu(x) = 0$
- Page 242, Line -4: Since Ω is compact **there exists $c_0 \geq 0$ such that $a_1 f_1 + \dots + a_i f_i + c_0 \geq 0$ on Ω for all $i = 1, \dots, \ell$.** We construct \dots
- Page 248, Line 1: \dots we get the far **left formula** in the proposition. To prove **the other two formulas**, we use \dots

- Page 282, Line 6: (ii) For $A, B \in \Sigma$ with $P(B) \neq 0$, $P(A|B) := \frac{P(A \cap B)}{P(B)}$...
- Page 283, Line 19:

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

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

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