Machine Learning

6-10-25
(1.te Einheit)

(1) Lineare Modelle – Lineare Diskriminanz Analyse (LDA)

Wir betrachten Modelle der Form

Klassen ID

$$(*) \quad p(y=c\,|\,x,\theta) = \frac{p(x\,|\,y=c,\theta)\cdot p(y=c)}{\sum_k p(x\,|\,y=k,\theta)\cdot p(y=k)}$$

Label/Target Daten Modellparameter

A-posteri Wahrscheinlichkeit der Klasse c, gegeben Datenpunkt x und fixen Modellparametern θ

$p(x\,|\,y=c,\theta)$ .... Klassenbedingte Wahrscheinlichkeitsdichte, gegeben Klasse c und fixen Modellparametern

$p(y=c)$ .... A-priori Wahrscheinlichkeit der Klasse c

Zu gegebenem x (Datenpunkt) und fixem θ, erhalten wir als Entscheidungsregel:

$$\boxed{k^* = \arg\max_c \; p(y=c\,|\,x,\theta)}$$

<u>Beispiel</u>: Wir betrachten den Fall $\underbrace{x \in \mathbb{R}^d}_{\text{Datenpunkt}}$ und $\underbrace{y \in \{0,1\}}_{\text{Label}}$.

<u>Annahme 1</u>: $p(x\,|\,y=c,\theta)$ lässt sich mittels einer multivariaten Normalverteilung beschreiben.

$$p(x \mid y=c, \theta_c) = (2\pi)^{-d/2} \cdot \sqrt{\det(\Sigma_c)} \cdot \exp\left(-\frac{1}{2} \cdot (x-\mu_c)^T \Sigma_c^{-1} (x-\mu_c)\right)$$

(MVN) ← Multivariate Normalverteilung mit <u>Kovarianz Matrix</u> $\underbrace{\Sigma_c}_{d \times d \text{ Matrix}}$ und Lokationsparameter $\mu_c \in \mathbb{R}^d$

mit $\theta_c = \{\Sigma_c, \mu_c\}$.

<u>Annahme 2</u>: $\underbrace{\Sigma_1 = \Sigma_0 = \Sigma}_{\text{Kovarianz Matrizen der Klasse 1 und 0 sind gleich!}}$

Wir sehen aus (*), dass

$$p(y=c \mid x, \theta) \underset{\text{proportional zu}}{\propto} p(x \mid y=c, \theta) \cdot p(y=c)$$

(da Nenner in (*) immer gleich)

Wir setzen $p(y=0 \mid x, \theta) = p(y=1 \mid x, \theta)$ um zu sehen wie die <u>Entscheidungsgrenze</u> aussieht. Definieren $\pi_0 = p(y=0)$ und $\pi_1 = p(y=1)$.

$$\pi_0 \cdot \cancel{\frac{1}{\sqrt{(2\pi)^d \cdot \det(\Sigma_0)}}} \cdot \exp\left(-\frac{1}{2} \cdot (x-\mu_0)^T \Sigma^{-1} (x-\mu_0)\right) = \cancel{\frac{1}{\sqrt{(2\pi)^d \cdot \det(\Sigma_1)}}} \cdot \exp\left(\ldots\ldots\right) \cdot \pi_1$$

$$\Sigma_0 = \Sigma, \quad \Sigma_1 = \Sigma$$

$$\overset{\log}{\Longleftrightarrow} \log(\pi_0) - \frac{1}{2} \cdot (x-\mu_0)^T \Sigma^{-1} (x-\mu_0) = \log(\pi_1) - \frac{1}{2} \cdot (x-\mu_1)^T \Sigma^{-1} (x-\mu_1)$$

$$\Longleftrightarrow \log(\pi_0) - \frac{1}{2} \cdot x^T \Sigma^{-1} x + \frac{1}{2} x^T \Sigma^{-1} \mu_0 + \frac{1}{2} \mu_0^T \Sigma^{-1} x - \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0 =$$

$$\log(\pi_1) - \frac{1}{2} x^T \Sigma^{-1} x + \frac{1}{2} \cdot x^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_1^T \Sigma^{-1} x - \frac{1}{2} \cdot \mu_1^T \Sigma^{-1} \mu_1$$

$$\Longleftrightarrow \cancel{-\frac{1}{2}x^T\Sigma^{-1}x} + \mu_0^T\Sigma^{-1}x - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \log(\pi_0) =$$
$$\cancel{-\frac{1}{2}x^T\Sigma^{-1}x} + \mu_1^T\Sigma^{-1}x - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log(\pi_1)$$

$$\Longleftrightarrow \left(\mu_1^T\Sigma^{-1}x - \mu_0^T\Sigma^{-1}x\right) - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \log\left(\frac{\pi_1}{\pi_0}\right) = 0$$

$$\Longleftrightarrow \left(\Sigma^{-1}(\mu_1-\mu_0)\right)^T x + \frac{1}{2}(\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0+\mu_1) + \log\left(\frac{\pi_1}{\pi_0}\right) = 0$$

$$\overset{\cdot 2}{\Longleftrightarrow} \underbrace{2\cdot\left(\Sigma^{-1}(\mu_1-\mu_0)\right)^T x}_{\text{abhängig von } x!!!} + \underbrace{(\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0+\mu_1)}_{\text{Skalar}} + \underbrace{2\cdot\log\left(\frac{\pi_1}{\pi_0}\right)}_{\text{Skalar}} = 0$$

check!!!

$\Longrightarrow$ Wir erhalten einen Ausdruck der Form $\boxed{a^Tx + b = 0}$ mit

$$a = \Sigma^{-1}(\mu_1-\mu_0)\cdot 2$$
$$b = (\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0+\mu_1) + 2\cdot\log\left(\frac{\pi_1}{\pi_0}\right)$$

<u>Lineare</u> Entscheidungs-grenze!

(daher der Name LDA)

$\downarrow$

Linear

<u>wichtig:</u> Wir haben $\Sigma_1 = \Sigma_0 = \Sigma$ angenommen. (Annahme 2).

Betrachten wir kurz

$$p(y=1\,|\,x,\theta) = \frac{p(x\,|\,y=1,\theta)\cdot \overbrace{p(y=1)}^{\pi_1}}{\underbrace{\text{Normalisierung}}_{\text{const.}}}$$

Einsetzen der MVN + Log., ergibt

$$\log p(y=1\,|\,x,\theta) = \log(\pi_1) - \frac{1}{2}\cdot(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \text{const.}$$

$$= -\frac{1}{2}\cdot\left[x^T\Sigma^{-1}x - 2\mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1\right] + \log(\pi_1) + \text{const.}$$

$$= \underbrace{\log(\pi_1) - \frac{1}{2}\cdot\mu_1^T\Sigma^{-1}\mu_1}_{\gamma_1} + \underbrace{\underbrace{\mu_1^T\Sigma^{-1}x}_{x^T\underbrace{\Sigma^{-1}\mu_1}_{\beta_1}}}_{x^T\beta_1} + \underbrace{\text{const} - \frac{1}{2}x^T\Sigma^{-1}x}_{\tau}$$

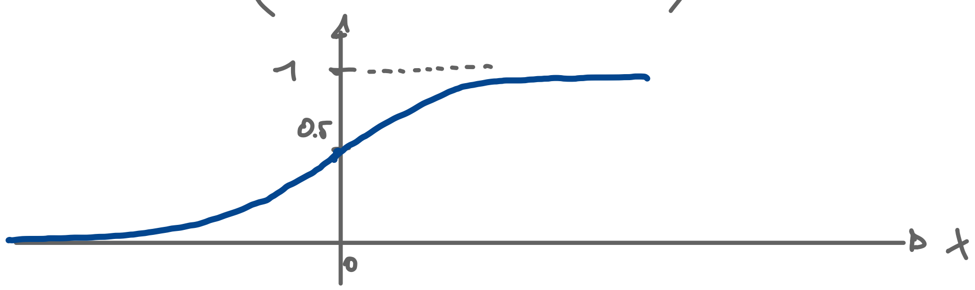
$$= \gamma_1 + x^T\beta_1 + \underbrace{\tau}_{\text{unabhängig von der Klasse}}$$

$$p(y=1\,|\,x,\theta) = \frac{\exp(x^T\beta_1 + \gamma_1)}{\exp(x^T\beta_0 + \gamma_0) + \exp(x^T\beta_1 + \gamma_1)}$$

$$= \frac{1}{1 + \exp\left((\beta_0-\beta_1)^T x + (\gamma_0 - \gamma_1)\right)}$$

__Warum?__ Mit $\sigma(x) := \frac{1}{1+\exp(-x)}$ (Sigmoid-Funktion) erhalten wir

$$p(y=1\,|\,x,\theta) = \sigma\left((\beta_1-\beta_0)^T x + (\gamma_1 - \gamma_0)\right) \qquad (*)$$



Betrachten wir $\nu_1 - \nu_0$:

$$\nu_1 - \nu_0 = \log(\pi_1) - \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 - \log(\pi_0) + \frac{1}{2}\cdot\mu_0^T\Sigma^{-1}\mu_0$$

$$= \log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}\cdot(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0)$$

Betrachten wir $\beta_1 - \beta_0$:

$$\beta_1 - \beta_0 = \Sigma^{-1}\mu_1 - \Sigma^{-1}\mu_0 = \Sigma^{-1}(\mu_1-\mu_0) =: w$$

<u>Ansatz:</u> $x_0 = \frac{1}{2}\cdot(\mu_1+\mu_0) - (\mu_1-\mu_0)\cdot\dfrac{\log\left(\frac{\pi_1}{\pi_0}\right)}{(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)}$

Warum? $w^T x_0 = (\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0)\cdot\frac{1}{2} - \cancel{(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)}\cdot\dfrac{\log\left(\frac{\pi_1}{\pi_0}\right)}{\cancel{(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)}}$

$\Rightarrow$ das ist genau $-(\nu_1-\nu_0)$ von oben.

$$w^T x_0 = -(\nu_1-\nu_0)$$

Einsetzen in $(*)$ <u>ergibt</u>:

$$p(y=1\,|\,x,\theta) = \sigma\big(\underbrace{(\beta_1-\beta_0)}_{w}{}^T x + (\nu_1-\nu_0)\big)$$

$$= \sigma(w^T x - w^T x_0)$$

$$= \sigma(w^T(x - x_0))$$

Datenpunkt

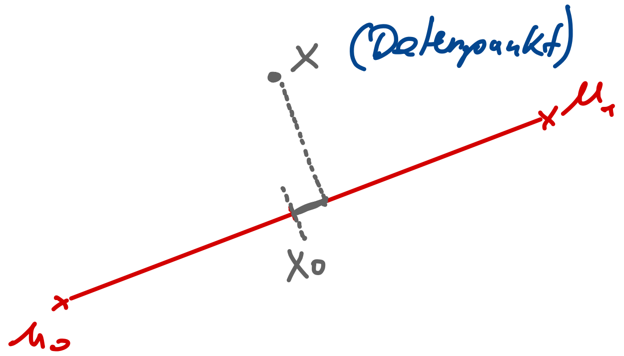
<u>Beobachtung 1:</u> Ist also $w^T x > w^T x_0$, also $w^T(x-x_0)$ positiv, dann ist $\sigma(w^T(x-x_0)) > 0.5 \Rightarrow p(y=1\,|\,x,\theta) > p(y=0\,|\,x,\theta)$

$\Rightarrow$ wir weisen x der Klasse 1 zu.

geometrisch: Wenn $\Sigma_1 = \Sigma_0 = \Sigma = \begin{pmatrix} 1 & & \emptyset \\ & \ddots & \\ \emptyset & & 1 \end{pmatrix}$ und $\pi_1 = \pi_0$, erhalten wir:

$$w = \Sigma^{-1}(\mu_1 - \mu_0), \text{ also } \mu_1 - \mu_0$$

$$x_0 = \frac{1}{2} \cdot (\mu_1 + \mu_0)$$



(Würden wir $\pi_1 > \pi_0$ haben (also nicht $\pi_1 = \pi_0$), dann hätten wir

$$p(y=1) > p(y=0)$$

und es reicht nicht mehr, über der "Mitte" zu liegen, um x der Klasse 0 zuzuweisen, sondern wir müssten näher an μ_0 rutschen.)

Einschub: Maximum-Likelihood-Estimation (MLE)

Daten $D = ((x_1, y_1) \ldots\ldots (x_N, y_N))$

$$\underbrace{p(y \mid x, \theta)}_{\text{Modell mit Parametern } \theta}$$

$\underbrace{p(D \mid \theta)}_{\text{W-keit der Daten } D \text{ unter Modell mit Parametern } \theta}$ .... Likelihood

Wir möchten θ so wählen, dass D die höchste Wahrscheinlichkeit hat (also den größten Likelihood).

$$\boxed{\theta_{mle} = \underset{\theta}{\operatorname{argmax}}\ p(\mathcal{D}\mid\theta)}$$

Annahme: Daten sind unabhängig und identisch verteilt (iid ... independent and identically distributed).

Likelihood

$$\theta_{mle} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p(y_i \mid x_i, \theta)$$

Log-Likelihood

$$\iff \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(y_i \mid x_i, \theta)$$

negative log-Likelihood (NLL)

$$\iff \underset{\theta}{\operatorname{argmin}} - \sum_{i=1}^{N} \log p(y_i \mid x_i, \theta)$$

Beispiel: Categorical (Cat) Distribution (kategorische Verteilung)

Diskrete Wahrscheinlichkeitsverteilung mit einem Parameter pro Klasse. (1, ..., C Klassen).

$$\boxed{\mathrm{Cat}(y\mid\theta) = \prod_{c=1}^{C} \theta_c^{\mathbb{1}_{y=c}}}$$

$$\mathbb{1}_{y=c} = \begin{cases} 1, & \text{wenn } y=c \\ 0, & \text{sonst} \end{cases}$$

also $p(y=c\mid\theta) = \theta_c \in [0,1]$, $\sum_c \theta_c = 1$

Andere Schreibweise: wir könnten $y=c$ auch als binär Vektor repräsentieren:

$$[0, 0, \dots 0, 1, 0 \dots]$$

↑ c-te Stelle (y_c)

Dann: $\mathrm{Cat}(y\mid\theta) = \prod_{c=1}^{C} \theta_c^{y_c}$

Für einen Datensatz $((x_1, y_1), \dots, (x_N, y_N)) = \mathcal{D}$ wollen wir den Likelihood unter unserem Modell anschreiben. Für $p(x \mid y=c, \theta)$ haben wir bereits die MVN mit Parametern (μ_c, Σ_c). Für $p(y=c)$ nehmen wir $\mathrm{Cat}(y \mid \pi)$.

π: a-priori Wahrscheinlichkeiten

Somit haben wir Parameter (θ) $\{\mu_1, \dots \mu_C, \Sigma_1 \dots \Sigma_C, \pi_1, \dots \pi_C\}$.

$$p(\mathcal{D} \mid \theta) = \prod_{n=1}^{N} \mathrm{Cat}(y_n \mid \pi) \cdot \prod_{c=1}^{C} \mathrm{MVN}(x_n \mid \mu_c, \Sigma_c)^{\mathbb{1}_{y_n=c}}$$

Logarithmieren:

$$\log p(\mathcal{D} \mid \theta) = \underbrace{\sum_{n=1}^{N} \sum_{c=1}^{C} \mathbb{1}_{y_n=c} \cdot \log(\pi_c)}_{} + \log\left(\prod_{n=1}^{N} \prod_{c=1}^{C} \mathrm{MVN}(x_n \mid \mu_c, \Sigma_c)^{\mathbb{1}_{y_n=c}}\right)$$

$$= \text{—}\,\text{n}\,\text{—} + \sum_{c=1}^{C} \sum_{n: y_n=c} \log \mathrm{MVN}(x_n \mid \mu_c, \Sigma_c)$$

Beobachtung: beide Terme und können separat maximiert werden (weil wir wollen ja $\underset{\theta}{\mathrm{argmax}}\, p(\mathcal{D} \mid \theta)$)

Term 1:

$$\sum_{n=1}^{N} \sum_{c=1}^{C} \mathbb{1}_{y_n=c} \cdot \log(\pi_c)$$

$$= \sum_{c=1}^{C} \sum_{n=1}^{N} \mathbb{1}_{y_n=c} \cdot \log(\pi_c)$$

$$= \sum_{c=1}^{C} \left[\log(\pi_c) \cdot \underbrace{\sum_{n=1}^{N} \mathbb{1}_{y_n=c}}_{=:N_c \text{ ... Anzahl an } y \text{ mit } y=c} \right]$$

$$= \sum_{c=1}^{C} N_c \cdot \log(\pi_c)$$

Wir versuchen $-\sum_{c=1}^{C} N_c \cdot \log(\pi_c)$ zu minimieren (unter der Nebenbedingung $\sum_c \pi_c = 1$).

Wir schreiben mit Hilfe von sog. Lagrange-Multiplizierern (λ)

$$-\sum_{c=1}^{C} N_c \cdot \log(\pi_c) - \lambda \cdot \underbrace{\left(1 - \sum_c \pi_c\right)}_{\text{Neben Bedingung}} = \mathcal{L}(\pi, \lambda)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -\left(1 - \sum_c \pi_c\right) = 0 \Rightarrow -1 + \sum_c \pi_c = 0 \Rightarrow \sum \pi_c = 1$$

$$\frac{\partial \mathcal{L}}{\partial \pi_c} = \underbrace{-\frac{N_c}{\pi_c} + \lambda}_{} = 0 \quad (\text{für } \pi_1, \dots, \pi_C)$$

$$\lambda = \frac{N_c}{\pi_c} \Rightarrow \lambda \pi_c = N_c \qquad \left(\text{wir wissen } \sum N_c = N\right)$$

d.h. $\sum_c \lambda \pi_c = \sum_c N_c = N \Rightarrow \boxed{\lambda = N}$

Wir wissen $\lambda \pi_c = N_c \Rightarrow \boxed{\pi_c = \frac{N_c}{N}}$

MLE für π

Term 2:

$$\underbrace{\sum_{c=1}^{C} \sum_{n:y_n=c} \log MVN(x_n \mid \mu_c, \Sigma_c)}$$

genau der <u>log-Likelihood</u> der MVN für Klasse c.

<u>Ohne Herleitung</u>

$$\boxed{\hat{\mu}_{c,mle} = \frac{1}{N_c} \cdot \sum_{n:y_n=c} x_n}$$

^ Notation = Schätzer

$$\boxed{\hat{\Sigma}_{c,mle} = \frac{1}{N_c} \cdot \sum_{n:y_n=c} (x_n - \mu_c) \cdot (x_n - \mu_c)^T}$$

Wenn wir wollen können wir $\underbrace{\Sigma_c = \Sigma}$ (also gleich für alle Klassen) setzen, und erhalten

→ in dem Fall, lin. Entscheidungsgrenzen (wie vorher)

$$\hat{\Sigma} = \frac{1}{N} \sum_{c=1}^{C} \sum_{n:y_n=c} (x_n - \hat{\mu}_c) \cdot (x_n - \hat{\mu}_c)^T$$

<u>Oder:</u> $\Sigma_c = \begin{pmatrix} \sigma_{1c}^2 & & \emptyset \\ & \sigma_{2c}^2 & \\ \emptyset & & \ddots \; \sigma_{dc}^2 \end{pmatrix}$ also eine Diagonalmatrix <u>pro Klasse</u>

Haben wir Σ_c <u>diagonal</u> und $\Sigma_c = \Sigma$ (also gleich für alle Klassen), nennen wir das "diagonal LDA".

<u>Anm.:</u> MVN mit μ und Σ:

$$\frac{1}{(2\pi)^{d/2} \cdot \sqrt{|\Sigma|}} \cdot \exp\left(-\frac{1}{2} \cdot (x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Mit Annahme von $\Sigma = \begin{pmatrix} \sigma_1^2 & & \emptyset \\ & \ddots & \\ \emptyset & & \sigma_d^2 \end{pmatrix}$ erhalten wir $\Sigma^{-1} = \begin{pmatrix} 1/\sigma_1^2 & & \emptyset \\ & \ddots & \\ \emptyset & & 1/\sigma_d^2 \end{pmatrix}$

und $|\Sigma| = \prod_{i=1}^{d} \sigma_i^2$. Weiters

$$-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) = \left(-\frac{1}{2}\right) \cdot \left[\sum_{i=1}^{d} (x_i - \mu_i)^2 \cdot \frac{1}{\sigma_i^2}\right]$$

$$\Rightarrow \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right) = \exp\left(-\frac{1}{2} \cdot \sum_{i=1}^{d} \frac{(x_i-\mu_i)^2}{\sigma_i^2}\right)$$

$$= \prod_{i=1}^{d} \exp\left(-\frac{1}{2} \frac{(x_i-\mu_i)^2}{\sigma_i^2}\right) \quad (a)$$

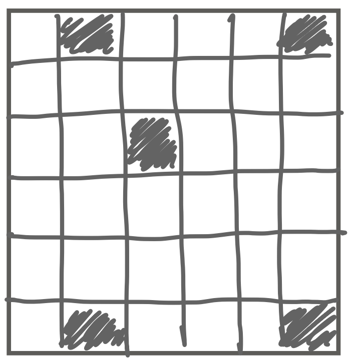
Bzgl. $\frac{1}{(2\pi)^{d/2} \cdot \sqrt{|\Sigma|}}$ erhalten wir:

$$\frac{1}{(2\pi)^{d/2} \sqrt{|\Sigma|}} = \frac{1}{(2\pi)^{d/2} \sqrt{\prod_{i=1}^{d} \sigma_i^2}} = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\, \sigma_i^2}} \quad (b)$$

Aus (a) und (b) erhalten wir schlussendlich

$$\underbrace{\prod_{i=1}^{d} \frac{1}{\sqrt{2\pi} \cdot \sigma_i} \cdot \exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right)}_{\text{Produkt von 1-D Normalverteilungen } \mathcal{N}(x_i \mid \mu_i, \sigma_i^2)}$$

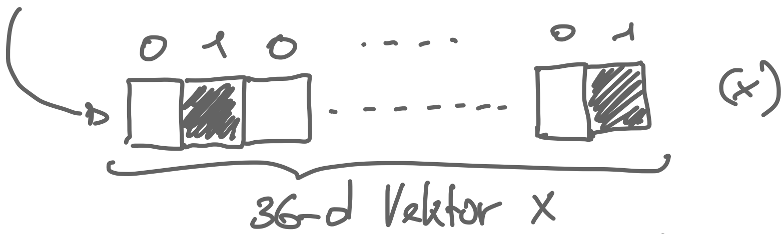
Beispiel (zu Normalverteilung als nicht-passendes Modell)



6x6 pixel "Bild"

Pixel sind entweder 0 od. 1 (also "Binärbild")

Vektorisieren



Hier passt die Normalverteilungsannahme nicht. Wir könnten aber jede einzelne Dimension in (x) als Bernoulli-verteilt modellieren, d.h.

$$\overbrace{\theta^{\mathbb{1}_{y=1}} \cdot (1-\theta)^{\mathbb{1}_{y=0}}}^{Ber(y|\theta)}$$

MLE (für Bernoulli Verteilung): Daten $\overbrace{y_1, \dots, y_N}^{D}$ $\quad y_i \in \{0,1\}$

$$p(D|\theta) = \prod_{i=1}^{N} \theta^{\mathbb{1}_{y_i=1}} \cdot (1-\theta)^{\mathbb{1}_{y_i=0}} \quad \text{// Likelihood}$$

$$\overset{\log}{\Longrightarrow} \log p(D|\theta) = \sum_{i=1}^{N} \left(\mathbb{1}_{y_i=1} \cdot \log(\theta) + \mathbb{1}_{y_i=0} \cdot \log(1-\theta) \right) \quad \text{// Log-Likelihood}$$

$$= \underline{N_1} \cdot \log(\theta) + \underline{N_0} \cdot \log(1-\theta)$$

N_0: Anzahl der y_i mit $y_i = 0$

N_1: Anzahl der y_i mit $y_i = 1$

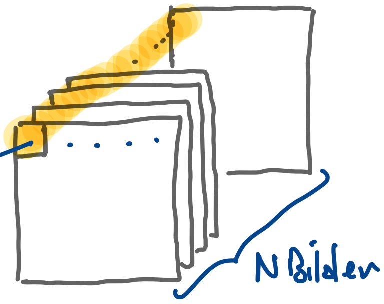
$\Rightarrow$ Negative Log-Likelihood (NLL): $NLL(\theta) = -\left(N_1 \cdot \log(\theta) + N_0 \cdot \log(1-\theta)\right)$

$$\frac{\partial NLL(\theta)}{\partial\theta} = -\frac{N_1}{\theta} + \frac{N_0}{1-\theta}$$

$\Rightarrow$ nach 0 setzen und auflösen nach θ erhalten wir:

$$\boxed{\hat{\theta}_{mle} = \frac{N_1}{N_0+N_1} = \frac{N_1}{N}}$$

D.h. in unserem Bild Beispiel:



1.ter Pixel

MLE zählt wie oft kommt '1' vor und dividiert durch N

Im Allgemeinen (also nicht auf Normalverteilung beschränkt) haben wir ein Modell der Form:

$$P(y=c\,|\,X,\theta) = \frac{P(y=c)\cdot\prod_{e=1}^{d} p(x_e\,|\,y=c,\theta_{ec})}{\sum_{c'=1}^{C} p(y=c')\cdot\prod_{e=1}^{d} p(x_e\,|\,y=c',\theta_{ec'})}$$

Daten $\in \mathbb{R}^d$

Wir nennen dies einen "Naive Bayes Classifier".

z.B. Parameter der Bernoulli-Verteilung für Koordinate e und Klasse c.

② Logistische Regression (LR)

Im Gegensatz zu LDA, ist LR ein $\underbrace{\text{diskriminativer}}$ Ansatz!

wir versuchen $p(y=c|x)$ <u>direkt</u> zu modellieren.

<u>1. Binärer Fall</u> $(y \in \{0,1\})$: Da $y \in \{0,1\}$, bietet es sich an $p(y|x)$ als Bernoulli-Verteilung zu modellieren, wobei wir den Parameter der Bernoulli-Verteilung abhängig vom Input (also unsere x) machen, z.B.

$$\mathrm{Ber}(y \,|\, \underbrace{f(x;\alpha)}_{\substack{\text{Funktion von } x \text{ mit Parameter } \alpha \\ (\text{wir wollen } f(x;\alpha) \in [0,1])}})$$

Nehmen wir an $x \in \mathbb{R}^d$, $y \in \{0,1\}$. In der LR haben wir folgendes Modell

$$p(y|x,\theta) = \mathrm{Ber}(y \,|\, \sigma(w^T x + b))$$

θ ↓ inkludiert jetzt w und b!

Setzen ab nun $a := w^T x + b$. Wir haben

$$p(y=1|x,\theta) = \sigma(a) = \frac{1}{1+e^{-a}}$$

$$p(y=0|x,\theta) = 1 - p(y=1|x,\theta) = 1 - \frac{1}{1+e^{-a}}$$

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

(Anm.: wir können uns b subsummiert in w denken, also
$w = [b, w_1, \dots, w_d]^T$ und $x = [1, x_1 \dots x_d]^T \Rightarrow$ in $\mathbb{R}^{d+1}$ eigentlich)

Wie bekommen wir w? MLE $\quad (x_1, y_1), \dots, (x_N, y_N)$; $x \in \mathbb{R}^d$, $y_n \in \{0, 1\}$

$$NLL(w) = -\frac{1}{N} \sum_{n=1}^{N} \log Ber(y_n | \underbrace{\sigma(w^T x_n)}_{=: \mu_n})$$

Einsetzen:

$$NLL(w) = -\frac{1}{N} \cdot \sum_{n=1}^{N} \log\left(\mu_n^{y_n} \cdot (1-\mu_n)^{1-y_n}\right)$$

$$= -\frac{1}{N} \cdot \sum_{n=1}^{N} \left[y_n \cdot \log(\mu_n) + (1-y_n) \cdot \log(1-\mu_n)\right]$$

$\mu_n \to \sigma(w^T x_n)$

Wir haben $w \in \mathbb{R}^d$ (od. $\mathbb{R}^{d+1}$). Wir versuchen den Gradienten von $NLL(w)$ (also $\nabla_w NLL(w)$) Null zu setzen und nach w auflösen.

$$\nabla_w NLL(w) = \begin{pmatrix} \frac{\partial NLL(w)}{\partial w_1} \\ \vdots \\ \frac{\partial NLL(w)}{\partial w_d} \end{pmatrix}$$

Wir wissen $\boxed{\begin{array}{l} \mu_n = \sigma(w^T x_n) \\ a_n = w^T x_n \end{array}}$ also $\mu_n = \sigma(a_n)$

1) $$\frac{\partial \mu_n}{\partial a_n} = \sigma(a_n) \cdot (1 - \sigma(a_n))$$

2) $$\frac{\partial \mu_n}{\partial w_d} = \frac{\partial}{\partial w_d} \sigma(\underbrace{w^T x_n}_{a_n}) = \underbrace{\frac{\partial}{\partial a_n} \sigma(a_n)}_{\text{siehe (1), also } \sigma(a_n) \cdot (1-\sigma(a_n))} \cdot \frac{\partial}{\partial w_d} a_n$$

$$\Rightarrow \frac{\partial \mu_n}{\partial w_d} = \sigma(a_n)\cdot\left(1-\sigma(a_n)\right)\cdot \underbrace{x_{nd}}_{\frac{\partial}{\partial w_d} a_n}$$

$$= \mu_n \cdot (1-\mu_n)\cdot x_{nd}$$

Wir haben also $\nabla_w \log(\mu_n) = \frac{1}{\mu_n}\cdot \nabla_w \mu_n$

$$= \begin{pmatrix} \frac{1}{\mu_n}\cdot\frac{\partial}{\partial w_1}\mu_n \\ \vdots \\ \frac{1}{\mu_n}\cdot\frac{\partial}{\partial w_d}\mu_n \end{pmatrix} = \begin{pmatrix} \cancel{\frac{1}{\mu_n}}\cdot\cancel{\mu_n}\cdot(1-\mu_n)\, x_{n1} \\ \vdots \\ \cancel{\frac{1}{\mu_n}}\cdot\cancel{\mu_n}\cdot(1-\mu_n)\cdot x_{nd} \end{pmatrix}$$

$$= \begin{pmatrix} (1-\mu_n)\cdot x_{n1} \\ \vdots \\ (1-\mu_n)\cdot x_{nd} \end{pmatrix} = \underbrace{(1-\mu_n)\cdot \begin{pmatrix} x_{n1} \\ \vdots \\ x_{nd} \end{pmatrix}}_{(1-\mu_n)\, x_n}$$

Auf gleiche Art u. Weise bekommen wir

$$\nabla_w \log(1-\mu_n) = -\mu_n \cdot \begin{pmatrix} x_{n1} \\ \vdots \\ x_{nd} \end{pmatrix} = -\mu_n \cdot x_n$$

$$\Rightarrow \nabla_w NLL(w) = -\frac{1}{N}\cdot \sum_{n=1}^{N} \left[y_n \cdot (1-\mu_n)\cdot x_n - (1-y_n)\,\mu_n \cdot x_n \right]$$

$$= -\frac{1}{N}\cdot \sum_{n=1}^{N} \left[y_n x_n - \cancel{y_n \mu_n x_n} - \mu_n x_n + \cancel{y_n \mu_n x_n} \right]$$

$$= +\frac{1}{N}\cdot \sum_{n=1}^{N} (\mu_n - y_n)\cdot x_n$$

x_n → Datenpunkt

y_n → Label $\in \{0,1\}$ (gewünscht)

μ_n → $\sigma(w^T x_n)$

Ziel ist ja $\nabla_w NLL(w) = 0$ nach w zu lösen.

Definieren

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & & \vdots \\ x_{N1} & \cdots & & x_{Nd} \end{pmatrix}$$

$N \times d$ Matrix
"Design Matrix"

$$\Rightarrow \nabla_w NLL(w) = \underbrace{X^T}_{d \times N} \underbrace{(\mu - y)}_{N \times 1} \cdot \frac{1}{N}$$

$$\mu = \begin{pmatrix} \sigma(w^T x_1) \\ \vdots \\ \sigma(w^T x_N) \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

Warum?

$$\frac{1}{N} \cdot \underbrace{\begin{pmatrix} x_{11} & \cdots & x_{N1} \\ \vdots & & \\ x_{1d} & \cdots & x_{Nd} \end{pmatrix}}_{X^T} \begin{pmatrix} \mu_1 - y_1 \\ \vdots \\ \mu_N - y_N \end{pmatrix} = \begin{pmatrix} x_{11} \cdot (\mu_1 - y_1) + x_{21} \cdot (\mu_2 - y_2) + \cdots \\ x_{12} \cdot (\mu_1 - y_1) + x_{22} \cdot (\mu_2 - y_2) + \cdots \\ \vdots \\ x_{1d} (\mu_1 - y_1) + \cdots \end{pmatrix}$$

x_1 $\qquad$ x_2

$$\boxed{\nabla_w NLL(w) = X^T (\mu - y) \cdot \frac{1}{N}}$$

Betrachten die Hesse-Matrix $H(w)$ $\longrightarrow$ $\underbrace{\begin{pmatrix} \frac{\partial NLL(\omega)}{\partial w_1 \partial w_1} & \cdots \\ \frac{\partial NLL(\omega)}{\partial w_1 \partial w_2} & \\ \vdots & \cdots \end{pmatrix}}_{d \times d \text{ Matrix}}$

Wir wissen

$$\frac{\partial}{\partial w_j} NLL(\omega) = \frac{1}{N} \cdot \sum_{n=1}^{N} (\underbrace{\mu_n}_{\sigma(w^T x_n)} - y_n) \cdot x_{nj}$$

$$\frac{\partial}{\partial w_j \partial w_k} NLL(\omega) = \frac{1}{N} \cdot \sum_{n=1}^{N} x_{nj} \frac{\partial}{\partial w_k} \mu_n$$ $\longrightarrow$ kennen wir schon von vorher

$$= \frac{1}{N} \cdot \sum_{n=1}^{N} x_{nj} \cdot x_{nk} \cdot (1-\mu_n) \cdot \mu_n$$

Mit $z_j = (x_{1j} \dots x_{Nj})^T$ und $z_k = (x_{1k} \dots x_{Nk})^T$, haben wir

$$\frac{\partial}{\partial w_j \partial w_k} NLL(\omega) = z_j^T B z_k$$

$$B = \underbrace{\begin{pmatrix} \mu_1 \cdot (1-\mu_1) & & \emptyset \\ & \ddots & \\ \emptyset & & \mu_N \cdot (1-\mu_N) \end{pmatrix}}_{N \times N \text{ Matrix}}$$

Also $\boxed{\begin{aligned} H(\omega) &= \nabla_w^2 NLL(w) \\ &= \frac{1}{N} \cdot X^T B X \end{aligned}}$

$\longrightarrow$ da B nur positive Einträge hat können wir B als $B^{\frac{1}{2}} \cdot B^{\frac{1}{2}}$ schreiben.

$$\Rightarrow H(\omega) = \frac{1}{N} \cdot X^T B^{\frac{1}{2}} \cdot B^{\frac{1}{2}} \cdot X$$

$$= \frac{1}{N} \cdot (B^{\frac{1}{2}} X)^T \cdot (B^{\frac{1}{2}} X)$$ also in der Form $A^T A$, $A = (B^{\frac{1}{2}} X)$

(Wir erinnern: eine reelle $N \times N$ Matrix C ist positiv semi-definit wenn für alle $b \in \mathbb{R}^N$ gilt:

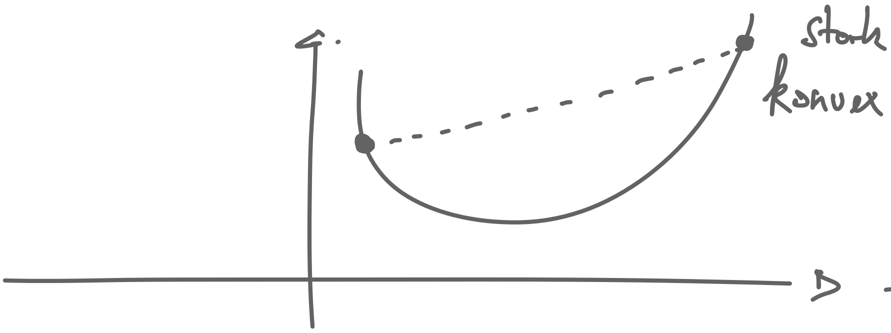
$$b^T C b \geq 0 \quad \text{(also nicht-negativ)}$$
)

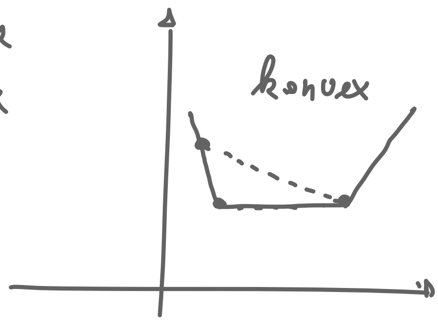
Bei uns: $b^T H(w)\, b = \left(b^T A^T A\, b\right) \cdot \frac{1}{N}$

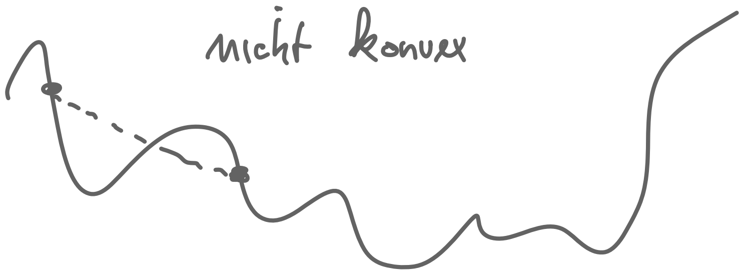
$$= (Ab)^T \cdot (Ab) \cdot \frac{1}{N} = \frac{1}{N} \cdot \|Ab\|^2 \geq 0 \quad \left(\text{für alle } b \in \mathbb{R}^N\right)$$

$\Rightarrow H(w)$ ist positiv semi-definit!

Eine 2-mal differenzierbare Funktion $f: \mathbb{R}^n \to \mathbb{R}$ ist konvex falls $\nabla^2 f$ positiv semi-definit ist.







Einschub: Stochastischer Gradientenabstieg (Stochastic Gradient Descent - SGD)