Machine Learning I

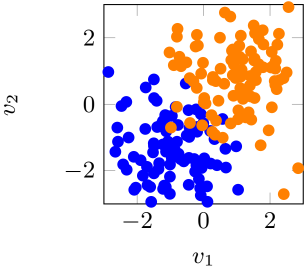
Bjoern Andres, Shengxian Zhao

Machine Learning for Computer Vision
TU Dresden

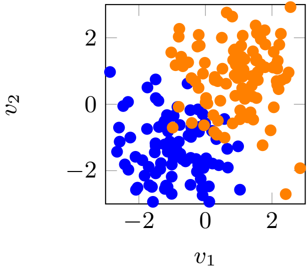
Winter Term 2022/2023

Contents. This part of the course is about a special case of supervised learning: the supervised learning of linear functions by **logistic regression**.

- We state the problem by defining labeled data, the family of functions and a **probability distribution** whose maximization motivates a regularizer and a loss function
- We show: This supervised learning problem is convex and can thus be solved by means of the **steepest descent algorithm**.

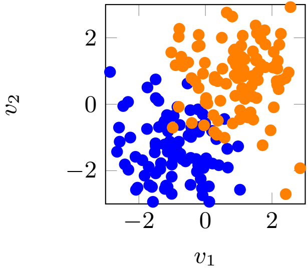


We consider **real attributes**. More specifically, we consider some finite set $V \neq \emptyset$ and labeled data $T = (S, X, x, y)$ with $X = \mathbb{R}^V$.



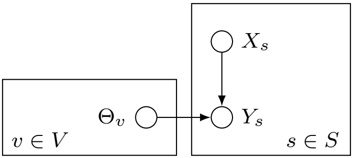
We consider **real attributes**. More specifically, we consider some finite set $V \neq \emptyset$ and labeled data $T = (S, X, x, y)$ with $X = \mathbb{R}^V$.

Hence, $x\colon S \to \mathbb{R}^V$ and $y\colon S \to \{0, 1\}$.



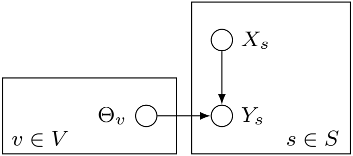
We consider **linear functions**. More specifically, we consider $\Theta = \mathbb{R}^V$ and $f : \Theta \to \mathbb{R}^X$ such that

$$\forall \theta \in \Theta \; \forall \hat{x} \in X : \quad f_\theta(\hat{x}) = \langle \theta, \hat{x} \rangle = \sum_{v \in V} \theta_v \, \hat{x}_v \tag{1}$$



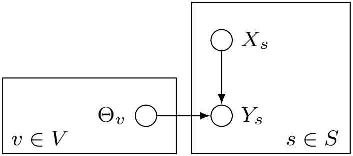
Random Variables

- For any sample $s \in S$, let X_s be a random variable whose value is a vector $x_s \in \mathbb{R}^V$, the **attribute vector** of s



Random Variables

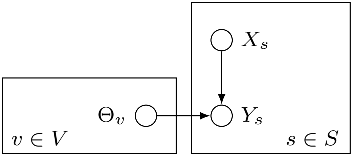
- For any sample $s \in S$, let X_s be a random variable whose value is a vector $x_s \in \mathbb{R}^V$, the **attribute vector** of s
- For any sample $s \in S$, let Y_s be a random variable whose value is a binary number $y_s \in \{0, 1\}$, the **label** of s



Random Variables

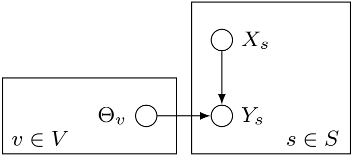
- For any sample $s \in S$, let X_s be a random variable whose value is a vector $x_s \in \mathbb{R}^V$, the **attribute vector** of s
- For any sample $s \in S$, let Y_s be a random variable whose value is a binary number $y_s \in \{0, 1\}$, the **label** of s
- For any $v \in V$, let Θ_v be a random variable whose value is a real number $\theta_v \in \mathbb{R}$, a **parameter** of the linear function we seek to learn

Deciding with Linear Functions

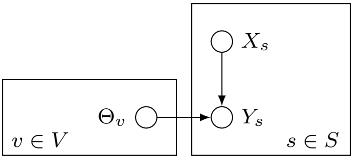


Factorization

$$P(X, Y, \Theta) = \prod_{s \in S} (P(Y_s \mid X_s, \Theta) P(X_s)) \prod_{v \in V} P(\Theta_v) \quad (2)$$

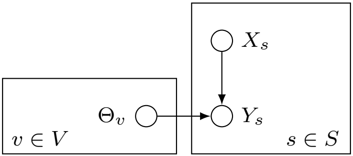


Factorization



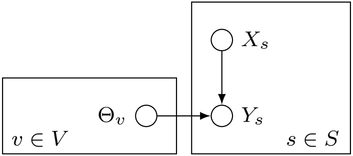
Factorization

$$P(\Theta \mid X, Y)$$



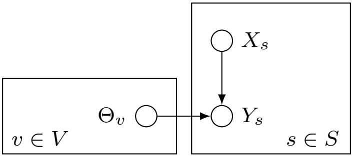
Factorization

$$P(\Theta \mid X, Y) = \frac{P(X, Y, \Theta)}{P(X, Y)}$$



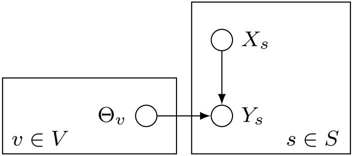
Factorization

$$
\begin{aligned}
P(\Theta \mid X, Y) &= \frac{P(X, Y, \Theta)}{P(X, Y)} \\
&= \frac{P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(X, Y)}
\end{aligned}
$$



Factorization

$$
\begin{aligned}
P(\Theta \mid X, Y) &= \frac{P(X, Y, \Theta)}{P(X, Y)} \\
&= \frac{P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(X, Y)} \\
&\propto P(Y \mid X, \Theta)\, P(\Theta)
\end{aligned}
$$



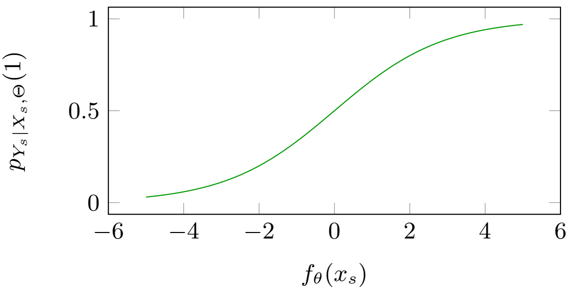
Factorization

$$\begin{aligned}
P(\Theta \mid X, Y) &= \frac{P(X, Y, \Theta)}{P(X, Y)} \\
&= \frac{P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(X, Y)} \\
&\propto P(Y \mid X, \Theta)\, P(\Theta) \\
&= \prod_{s \in S} P(Y_s \mid X_s, \Theta) \prod_{v \in V} P(\Theta_v)
\end{aligned}$$

Distributions

- **Logistic distribution**

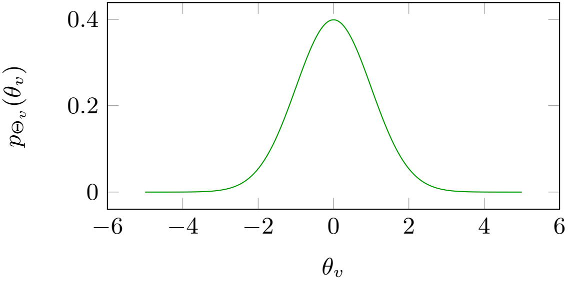
$$\forall s \in S : \qquad p_{Y_s \mid X_s, \Theta}(1) = \frac{1}{1 + 2^{-f_\theta(x_s)}} \tag{3}$$



Distributions

- **Normal distribution** with $\sigma \in \mathbb{R}^+$:

$$\forall v \in V : \qquad p_{\Theta_v}(\theta_v) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\theta_v^2/2\sigma^2} \tag{3}$$



Lemma. Estimating maximally probable parameters θ, given attributes x and labels y, i.e.,

$$\underset{\theta \in \mathbb{R}^m}{\operatorname{argmax}} \quad p_{\Theta|X,Y}(\theta, x, y)$$

is equivalent ot the supervised learning problem

$$\min_{\theta \in \Theta} \quad \lambda R(\theta) + \frac{1}{|S|} \sum_{s \in S} L(f_\theta(x_s), y_s) \tag{4}$$

with L, R and λ such that

$$\forall r \in \mathbb{R} \; \forall \hat{y} \in \{0,1\}\colon \quad L(r, \hat{y}) = -\hat{y}r + \log\left(1 + 2^r\right) \tag{5}$$

$$\forall \theta \in \Theta\colon \quad R(\theta) = \|\theta\|_2^2 \tag{6}$$

$$\lambda = \frac{\log e}{2\sigma^2} \quad . \tag{7}$$

Lemma. Estimating maximally probable parameters θ, given attributes x and labels y, i.e.,

$$\underset{\theta \in \mathbb{R}^m}{\operatorname{argmax}} \quad p_{\Theta|X,Y}(\theta, x, y)$$

is equivalent ot the supervised learning problem

$$\min_{\theta \in \Theta} \quad \lambda R(\theta) + \frac{1}{|S|} \sum_{s \in S} L(f_\theta(x_s), y_s) \tag{4}$$

with L, R and λ such that

$$\forall r \in \mathbb{R}\ \forall \hat{y} \in \{0,1\}\colon \quad L(r, \hat{y}) = -\hat{y}r + \log\left(1 + 2^r\right) \tag{5}$$

$$\forall \theta \in \Theta\colon \quad R(\theta) = \|\theta\|_2^2 \tag{6}$$

$$\lambda = \frac{\log e}{2\sigma^2} \quad . \tag{7}$$

It is called the l_2-regularized **logistic regression problem** with respect to x, y and σ.

Proof. Firstly,

$$
\begin{aligned}
& \underset{\theta \in \mathbb{R}^m}{\operatorname{argmax}} \quad p_{\Theta \mid X,Y}(\theta, x, y) \\
= & \underset{\theta \in \mathbb{R}^m}{\operatorname{argmax}} \quad \prod_{s \in S} p_{Y_s \mid X_s, \Theta}(y_s, x_s, \theta) \prod_{v \in V} p_{\Theta_v}(\theta_v) \\
= & \underset{\theta \in \mathbb{R}^m}{\operatorname{argmax}} \quad \sum_{s \in S} \log p_{Y_s \mid X_s, \Theta}(y_s, x_s, \theta) + \sum_{v \in V} \log p_{\Theta_v}(\theta_v) \qquad (8)
\end{aligned}
$$

Proof. Firstly,

$$
\begin{aligned}
& \underset{\theta \in \mathbb{R}^m}{\operatorname{argmax}} \quad p_{\Theta|X,Y}(\theta, x, y) \\
= & \underset{\theta \in \mathbb{R}^m}{\operatorname{argmax}} \quad \prod_{s \in S} p_{Y_s|X_s,\Theta}(y_s, x_s, \theta) \prod_{v \in V} p_{\Theta_v}(\theta_v) \\
= & \underset{\theta \in \mathbb{R}^m}{\operatorname{argmax}} \quad \sum_{s \in S} \log p_{Y_s|X_s,\Theta}(y_s, x_s, \theta) + \sum_{v \in V} \log p_{\Theta_v}(\theta_v)
\end{aligned}
\tag{8}
$$

Secondly,

$$
\begin{aligned}
& \log p_{Y_s|X_s,\Theta}(y_s, x_s, \theta) \\
= \quad & y_s \log p_{Y_s|X_s,\Theta}(1, x_s, \theta) + (1 - y_s) \log p_{Y_s|X_s,\Theta}(0, x_s, \theta) \\
= \quad & y_s \log \frac{p_{Y_s|X_s,\Theta}(1, x_s, \theta)}{p_{Y_s|X_s,\Theta}(0, x_s, \theta)} + \log p_{Y_s|X_s,\Theta}(0, x_s, \theta)
\end{aligned}
\tag{9}
$$

Proof. Firstly,

$$\begin{aligned} & \underset{\theta \in \mathbb{R}^m}{\operatorname{argmax}} \quad p_{\Theta|X,Y}(\theta, x, y) \\ = & \underset{\theta \in \mathbb{R}^m}{\operatorname{argmax}} \quad \prod_{s \in S} p_{Y_s|X_s,\Theta}(y_s, x_s, \theta) \prod_{v \in V} p_{\Theta_v}(\theta_v) \\ = & \underset{\theta \in \mathbb{R}^m}{\operatorname{argmax}} \quad \sum_{s \in S} \log p_{Y_s|X_s,\Theta}(y_s, x_s, \theta) + \sum_{v \in V} \log p_{\Theta_v}(\theta_v) \end{aligned} \tag{8}$$

Secondly,

$$\begin{aligned} & \log p_{Y_s|X_s,\Theta}(y_s, x_s, \theta) \\ = \quad & y_s \log p_{Y_s|X_s,\Theta}(1, x_s, \theta) + (1 - y_s) \log p_{Y_s|X_s,\Theta}(0, x_s, \theta) \\ = \quad & y_s \log \frac{p_{Y_s|X_s,\Theta}(1, x_s, \theta)}{p_{Y_s|X_s,\Theta}(0, x_s, \theta)} + \log p_{Y_s|X_s,\Theta}(0, x_s, \theta) \end{aligned} \tag{9}$$

Thus, with (3) and (4):

$$\underset{\theta \in \mathbb{R}^m}{\operatorname{argmin}} \quad \sum_{s \in S} \left(-y_s \langle \theta, x_s \rangle + \log \left(1 + 2^{\langle \theta, x_s \rangle} \right) \right) + \frac{\log e}{2\sigma^2} \|\theta\|_2^2 \tag{10}$$

Lemma. The objective function

$$\varphi(\theta) \;=\; \lambda R(\theta) + \frac{1}{|S|} \sum_{s \in S} L(f_\theta(x_s), y_s) \tag{11}$$

of the l_2-regularized logistic regression problem is convex.

Lemma. The objective function

$$\varphi(\theta) \;=\; \lambda R(\theta) + \frac{1}{|S|}\sum_{s\in S} L(f_\theta(x_s), y_s) \tag{11}$$

of the l_2-regularized logistic regression problem is convex.

Proof. Exercise!

Lemma. The objective function

$$\varphi(\theta) \;=\; \lambda R(\theta) + \frac{1}{|S|} \sum_{s \in S} L(f_\theta(x_s), y_s) \tag{11}$$

of the l_2-regularized logistic regression problem is convex.

Proof. Exercise!

The problem can be solved by the steepest descent algorithm with a tolerance parameter $\epsilon \in \mathbb{R}_0^+$:

$\theta := 0$
repeat
    $d := \nabla \varphi(\theta)$
    $\eta := \operatorname{argmin}_{\eta' \in \mathbb{R}} \varphi(\theta - \eta' d)$ (line search)
    $\theta := \theta - \eta d$
    if $\|d\| < \epsilon$
        return θ

Lemma: Estimating maximally probable labels y, given attributes x' and parameters θ, i.e.,

$$\underset{y \in \{0,1\}^S}{\operatorname{argmax}} \quad p_{Y|X,\Theta}(y, x', \theta) \tag{12}$$

is equivalent to the inference problem

$$\min_{y' \in \{0,1\}^S} \sum_{s \in S} L(f_\theta(x_s), y'_s) \ . \tag{13}$$

It has the solution

$$\forall s \in S' : \quad y_s = \begin{cases} 1 & \text{if } f_\theta(x'_s) > 0 \\ 0 & \text{otherwise} \end{cases} \ . \tag{14}$$

Proof. Firstly,

$$\underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad p_{Y|X,\Theta}(y, x', \theta)$$

Proof. Firstly,

$$\begin{aligned} & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad p_{Y|X,\Theta}(y, x', \theta) \\ = & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \prod_{s \in S'} p_{Y_s|X_s,\Theta}(y_s, x'_s, \theta) \end{aligned}$$

Proof. Firstly,

$$
\begin{aligned}
& \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad p_{Y|X,\Theta}(y, x', \theta) \\
= & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \prod_{s \in S'} p_{Y_s|X_s,\Theta}(y_s, x'_s, \theta) \\
= & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \sum_{s \in S'} \log p_{Y_s|X_s,\Theta}(y_s, x'_s, \theta)
\end{aligned}
$$

Proof. Firstly,

$$
\begin{aligned}
& \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad p_{Y|X,\Theta}(y, x', \theta) \\
= & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \prod_{s \in S'} p_{Y_s|X_s,\Theta}(y_s, x'_s, \theta) \\
= & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \sum_{s \in S'} \log p_{Y_s|X_s,\Theta}(y_s, x'_s, \theta) \\
= & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \sum_{s \in S'} \left(y_s \log \frac{p_{Y_s|X_s,\Theta}(1, x'_s, \theta)}{p_{Y_s|X_s,\Theta}(0, x'_s, \theta)} + \log p_{Y_s|X_s,\Theta}(0, x'_s, \theta) \right)
\end{aligned}
$$

Proof. Firstly,

$$
\begin{aligned}
& \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad p_{Y|X,\Theta}(y, x', \theta) \\
= & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \prod_{s \in S'} p_{Y_s|X_s,\Theta}(y_s, x'_s, \theta) \\
= & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \sum_{s \in S'} \log p_{Y_s|X_s,\Theta}(y_s, x'_s, \theta) \\
= & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \sum_{s \in S'} \left(y_s \log \frac{p_{Y_s|X_s,\Theta}(1, x'_s, \theta)}{p_{Y_s|X_s,\Theta}(0, x'_s, \theta)} + \log p_{Y_s|X_s,\Theta}(0, x'_s, \theta) \right) \\
= & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmin}} \quad \sum_{s \in S'} \left(-y_s f_\theta(x'_s) + \log \left(1 + 2^{f_\theta(x'_s)} \right) \right)
\end{aligned}
$$

Proof. Firstly,

$$
\begin{aligned}
& \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad p_{Y|X,\Theta}(y, x', \theta) \\
= \; & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \prod_{s \in S'} p_{Y_s|X_s,\Theta}(y_s, x'_s, \theta) \\
= \; & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \sum_{s \in S'} \log p_{Y_s|X_s,\Theta}(y_s, x'_s, \theta) \\
= \; & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \sum_{s \in S'} \left(y_s \log \frac{p_{Y_s|X_s,\Theta}(1, x'_s, \theta)}{p_{Y_s|X_s,\Theta}(0, x'_s, \theta)} + \log p_{Y_s|X_s,\Theta}(0, x'_s, \theta) \right) \\
= \; & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmin}} \quad \sum_{s \in S'} \left(-y_s f_\theta(x'_s) + \log \left(1 + 2^{f_\theta(x'_s)} \right) \right) \\
= \; & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmin}} \quad \sum_{s \in S'} L(f_\theta(x'_s), y_s) \quad .
\end{aligned}
$$

Proof. Firstly,

$$
\begin{aligned}
& \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad p_{Y|X,\Theta}(y, x', \theta) \\
= & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \prod_{s \in S'} p_{Y_s|X_s,\Theta}(y_s, x'_s, \theta) \\
= & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \sum_{s \in S'} \log p_{Y_s|X_s,\Theta}(y_s, x'_s, \theta) \\
= & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmax}} \quad \sum_{s \in S'} \left(y_s \log \frac{p_{Y_s|X_s,\Theta}(1, x'_s, \theta)}{p_{Y_s|X_s,\Theta}(0, x'_s, \theta)} + \log p_{Y_s|X_s,\Theta}(0, x'_s, \theta) \right) \\
= & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmin}} \quad \sum_{s \in S'} \left(-y_s f_\theta(x'_s) + \log \left(1 + 2^{f_\theta(x'_s)} \right) \right) \\
= & \underset{y \in \{0,1\}^{S'}}{\operatorname{argmin}} \quad \sum_{s \in S'} L(f_\theta(x'_s), y_s) \quad .
\end{aligned}
$$

Secondly,

$$
\min_{y \in \{0,1\}^{S'}} \sum_{s \in S'} \left(-y_s f_\theta(x'_s) + \log \left(1 + 2^{f_\theta(x'_s)} \right) \right)
$$

Proof. Firstly,

$$\begin{aligned}
&\underset{y\in\{0,1\}^{S'}}{\operatorname{argmax}} \quad p_{Y|X,\Theta}(y,x',\theta) \\
=\ &\underset{y\in\{0,1\}^{S'}}{\operatorname{argmax}} \quad \prod_{s\in S'} p_{Y_s|X_s,\Theta}(y_s,x'_s,\theta) \\
=\ &\underset{y\in\{0,1\}^{S'}}{\operatorname{argmax}} \quad \sum_{s\in S'} \log p_{Y_s|X_s,\Theta}(y_s,x'_s,\theta) \\
=\ &\underset{y\in\{0,1\}^{S'}}{\operatorname{argmax}} \quad \sum_{s\in S'} \left(y_s \log \frac{p_{Y_s|X_s,\Theta}(1,x'_s,\theta)}{p_{Y_s|X_s,\Theta}(0,x'_s,\theta)} + \log p_{Y_s|X_s,\Theta}(0,x'_s,\theta) \right) \\
=\ &\underset{y\in\{0,1\}^{S'}}{\operatorname{argmin}} \quad \sum_{s\in S'} \left(-y_s f_\theta(x'_s) + \log\left(1+2^{f_\theta(x'_s)}\right)\right) \\
=\ &\underset{y\in\{0,1\}^{S'}}{\operatorname{argmin}} \quad \sum_{s\in S'} L(f_\theta(x'_s),y_s) \ .
\end{aligned}$$

Secondly,

$$\min_{y\in\{0,1\}^{S'}} \sum_{s\in S'} \left(-y_s f_\theta(x'_s) + \log\left(1+2^{f_\theta(x'_s)}\right)\right) = \sum_{s\in S'} \max_{y_s\in\{0,1\}} y_s f_\theta(x'_s) \ .$$

Summary.

- The l_2-regularized logistic regression problem is a supervised learning problem w.r.t. the family of linear functions.
- It is motivated by a Bayesian statistical model with the logistic distribution as the likelihood as the normal distribution as the prior.
- It is a convex optimization problem that can be solved, e.g., by the steepest descent algorithm.