

Machine Learning I

Bjoern Andres

Machine Learning for Computer Vision
TU Dresden



<https://mlcv.cs.tu-dresden.de/courses/26-winter/ml1/>

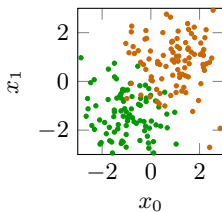
Winter Term 2026/2027

Supervised Learning of Linear Functions

Contents. This part of the course is about the supervised learning of linear functions, more specifically, about linear logistic regression and linear support vector machines.

- ▶ We introduce these problems by defining labeled data, a family of functions, a regularizer and a loss function.
- ▶ We motivate the logistic loss function probabilistically.
- ▶ We show: l_2 -regularized linear logistic regression is a convex optimization problem. It can be solved, e.g., by the steepest descent algorithm with line search.
- ▶ We state the l_1 -regularized linear support vector machine learning problem in the form of a linear program.

Supervised Learning of Linear Functions

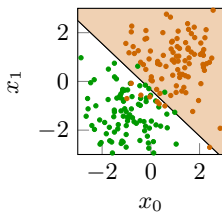


We consider labeled data with **real inputs**. More specifically, we consider some finite, non-empty set V , called the set of inputs, and labeled data $T = (S, X, x, y)$ such that $X = \mathbb{R}^V$. Hence:

$$x: S \rightarrow \mathbb{R}^V \tag{1}$$

$$y: S \rightarrow \{0, 1\} \tag{2}$$

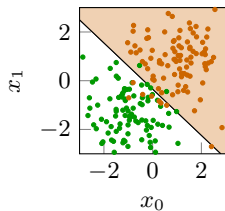
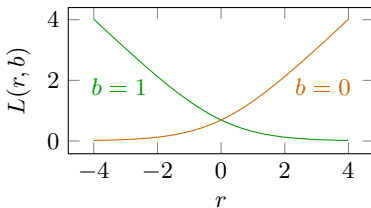
Supervised Learning of Linear Functions



We consider **linear functions**. More specifically, we consider $\Theta = \mathbb{R}^V$ and $f : \Theta \rightarrow \mathbb{R}^X$ such that for all $\theta \in \Theta$ and all $\hat{x} \in X$:

$$f_{\theta}(\hat{x}) = \langle \theta, \hat{x} \rangle = \sum_{v \in V} \theta_v \hat{x}_v . \quad (3)$$

Supervised Learning of Linear Functions



Example: For

- ▶ any labeled data (S, X, x, y) such that $X = \mathbb{R}^V$ and $V \neq \emptyset$ finite
- ▶ $\Theta = \mathbb{R}^V$ and $f: \Theta \rightarrow \mathbb{R}^X$ such that $f_\theta(\cdot) = \langle \theta, \cdot \rangle$
- ▶ $L: \mathbb{R} \times \{0, 1\} \rightarrow \mathbb{R}_0^+$ such that for any $r \in \mathbb{R}$ and any $b \in \{0, 1\}$:

$$L(r, b) = -rb + \ln(1 + e^r) = \ln(1 + e^{-(2b-1)r}) \quad (4)$$

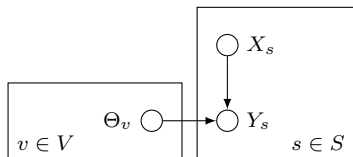
- ▶ any $R: \Theta \rightarrow \mathbb{R}_0^+$ and any $\lambda \in \mathbb{R}_0^+$

the supervised learning problem

$$\inf_{\theta \in \Theta} \lambda R(\theta) + \sum_{s \in S} L(f_\theta(x_s), y_s) \quad (5)$$

is called the **linear logistic regression problem**.

Supervised Learning of Linear Functions



We motivate the logistic loss function probabilistically:

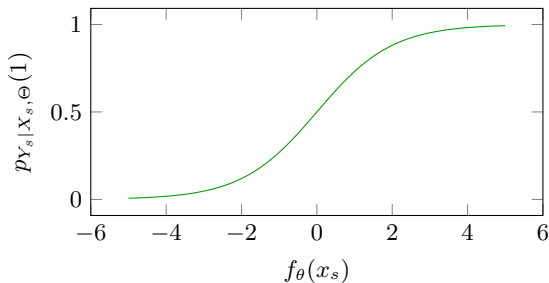
- For any sample $s \in S$, let X_s be a random variable whose value is a vector $x_s \in \mathbb{R}^V$, the **input vector** of s
- For any sample $s \in S$, let Y_s be a random variable whose value is a binary number $y_s \in \{0, 1\}$, the **label** of s
- For any $v \in V$, let Θ_v be a random variable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

We assume the conditional independencies of the Bayes net depicted above.
Thus,

$$P(X, Y, \Theta) = \prod_{s \in S} (P(Y_s \mid X_s, \Theta) P(X_s)) \prod_{v \in V} P(\Theta_v) . \quad (6)$$

► Sigmoid distribution

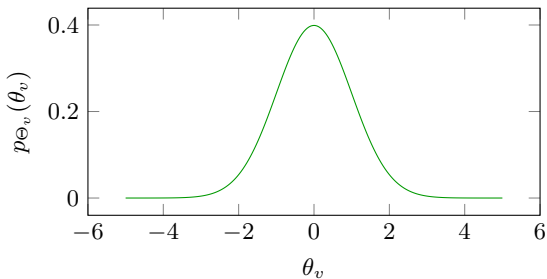
$$\forall s \in S : \quad p_{Y_s|X_s, \Theta}(1) = \frac{1}{1 + e^{-f_{\theta}(x_s)}} \quad (7)$$



Supervised Learning of Linear Functions

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

$$\forall v \in V : \quad p_{\Theta_v}(\theta_v) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\theta_v^2/2\sigma^2} \quad (8)$$



Supervised Learning of Linear Functions

We attempt to **learn** parameters θ from given inputs x and given labels y by maximizing wrt. θ :

$$p_{\Theta|X,Y}(\theta, x, y) = \frac{p_{X,Y,\Theta}(x, y, \theta)}{p_{X,Y}(x, y)} \quad (9)$$

$$= \frac{p_{Y|X,\Theta}(y, x, \theta) p_X(x) p_{\Theta}(\theta)}{p_{X,Y}(x, y)} \quad (10)$$

$$\propto p_{Y|X,\Theta}(y, x, \theta) p_{\Theta}(\theta) \quad (11)$$

$$= \prod_{s \in S} p_{Y_s|X_s,\Theta}(y_s, x_s, \theta) \prod_{v \in V} p_{\Theta_v}(\theta_v) . \quad (12)$$

We attempt to **infer** labels from given inputs x and given parameters θ by maximizing wrt. y :

$$p_{Y|X,\Theta}(y) . \quad (13)$$

Supervised Learning of Linear Functions

Lemma. The problem of learning the parameters θ for given inputs x and given labels y , i.e.,

$$\operatorname{argmax}_{\theta \in \mathbb{R}^V} \prod_{s \in S} p_{Y_s | X_s, \Theta}(\theta) \prod_{v \in V} p_v(\theta_v)$$

is equivalent to the l_2 -regularized logistic regression problem, more specifically, to the supervised learning problem

$$\operatorname{argmin}_{\theta \in \Theta} \quad \lambda R(\theta) + \sum_{s \in S} L(f_\theta(x_s), y_s) \quad (14)$$

with L , R and λ such that

$$\forall r \in \mathbb{R} \quad \forall b \in \{0, 1\}: \quad L(r, b) = -rb + \ln(1 + e^r) \quad (15)$$

$$\forall \theta \in \Theta: \quad R(\theta) = \|\theta\|_2^2 \quad (16)$$

$$\lambda = \frac{1}{2\sigma^2} \quad . \quad (17)$$

Supervised Learning of Linear Functions

Proof. Firstly,

$$\begin{aligned} & \operatorname{argmax}_{\theta \in \mathbb{R}^V} \prod_{s \in S} p_{Y_s|X_s, \Theta}(y_s, x_s, \theta) \prod_{v \in V} p_{\Theta_v}(\theta_v) \\ &= \operatorname{argmax}_{\theta \in \mathbb{R}^V} \sum_{s \in S} \ln p_{Y_s|X_s, \Theta}(y_s, x_s, \theta) + \sum_{v \in V} \ln p_{\Theta_v}(\theta_v) \end{aligned} \quad (18)$$

Secondly,

$$\begin{aligned} & \ln p_{Y_s|X_s, \Theta}(y_s, x_s, \theta) \\ &= y_s \ln p_{Y_s|X_s, \Theta}(1, x_s, \theta) + (1 - y_s) \ln p_{Y_s|X_s, \Theta}(0, x_s, \theta) \\ &= y_s \ln \frac{p_{Y_s|X_s, \Theta}(1, x_s, \theta)}{p_{Y_s|X_s, \Theta}(0, x_s, \theta)} + \ln p_{Y_s|X_s, \Theta}(0, x_s, \theta) \end{aligned} \quad (19)$$

Thus, with (7) and (8):

$$\operatorname{argmin}_{\theta \in \mathbb{R}^V} \sum_{s \in S} \left(-y_s f_{\theta}(x_s) + \ln \left(1 + e^{f_{\theta}(x_s)} \right) \right) + \frac{1}{2\sigma^2} \|\theta\|_2^2 \quad (20)$$

Lemma. For the linear function $f_\theta(\cdot) = \langle \theta, \cdot \rangle$, the objective function

$$\varphi(\theta) = \lambda R(\theta) + \sum_{s \in S} L(f_\theta(x_s), y_s) \quad (21)$$

of the l_2 -regularized logistic regression problem is convex.

Proof. Exercise!

The l_2 -regularized linear logistic regression problem can be solved, e.g., by the steepest descent algorithm with line search:

Algorithm. Steepest descent with line search and a 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  $\theta$ 
```

Lemma. The problem of inferring maximally probable labels y , given inputs x and parameters θ , i.e.,

$$\operatorname{argmax}_{y \in \{0,1\}^S} p_{Y|X,\Theta}(y) \quad (22)$$

has the same solutions as the inference problem

$$\operatorname{argmin}_{y \in \{0,1\}^S} \sum_{s \in S} L(f_{\theta}(x_s), y_s) . \quad (23)$$

These are such that for all $s \in S$:

$$y_s = \begin{cases} 0 & \text{if } f_{\theta}(x_s) < 0 \\ 1 & \text{if } f_{\theta}(x_s) > 0 \end{cases} . \quad (24)$$

Supervised Learning of Linear Functions

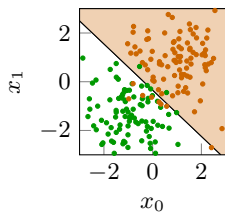
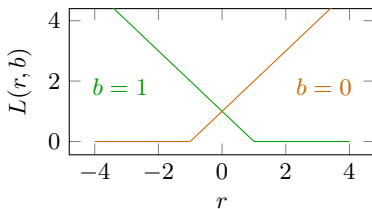
Proof. Firstly,

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

Secondly,

$$\min_{y \in \{0,1\}^S} \sum_{s \in S} \left(-y_s f_\theta(x_s) + \ln \left(1 + e^{f_\theta(x_s)} \right) \right) = \sum_{s \in S} \max_{y_s \in \{0,1\}} y_s f_\theta(x_s) .$$

Supervised Learning of Linear Functions



Example: For

- ▶ any labeled data (S, X, x, y) such that $X = \mathbb{R}^V$ and $V \neq \emptyset$ finite
- ▶ $\Theta = \mathbb{R}^V$ and $f: \Theta \rightarrow \mathbb{R}^X$ such that $f_\theta(\cdot) = \langle \theta, \cdot \rangle$
- ▶ $L: \mathbb{R} \times \{0, 1\} \rightarrow \mathbb{R}_0^+$ such that for any $r \in \mathbb{R}$ and any $b \in \{0, 1\}$:

$$L(r, b) = \max(0, -(2b - 1)r + 1) \quad (\text{hinge loss function}) \quad (25)$$

- ▶ any $R: \Theta \rightarrow \mathbb{R}_0^+$ and any $\lambda \in \mathbb{R}_0^+$,

the supervised learning problem

$$\inf_{\theta \in \Theta} \lambda R(\theta) + \sum_{s \in S} L(f_\theta(x_s), y_s) \quad (26)$$

is called the **linear support vector machine (SVM) learning problem**.

Supervised Learning of Linear Functions

Lemma: For

- ▶ any labeled data (S, X, x, y) such that $X = \mathbb{R}^V$ and $V \neq \emptyset$ finite,
- ▶ $\Theta = \mathbb{R}^V$ and $f: \Theta \rightarrow \mathbb{R}^X$ such that $f_\theta(\cdot) = \langle \theta, \cdot \rangle$
- ▶ $R: \Theta \rightarrow \mathbb{R}_0^+$ such that $R(\cdot) = \|\cdot\|_1$,
- ▶ any $\lambda \in \mathbb{R}_0^+$, and
- ▶ any $\hat{\theta} \in \Theta$,

$\hat{\theta}$ is a solution to the linear SVM learning problem wrt. the above iff there exist $\hat{\xi} \in \mathbb{R}^S$ and $\hat{\theta}^+, \hat{\theta}^- \in \mathbb{R}^V$ such that $\hat{\theta} = \hat{\theta}^+ - \hat{\theta}^-$ and $(\hat{\xi}, \hat{\theta}^+, \hat{\theta}^-)$ is a solution to the linear program

$$\min_{\theta \in \Theta, \xi \in \mathbb{R}^S} \sum_{s \in S} \xi_s + \lambda \sum_{v \in V} \theta_v^+ + \lambda \sum_{v \in V} \theta_v^- \quad (27)$$

$$\text{subject to } \forall s \in S: (2y_s - 1)\langle \theta^+ - \theta^-, x_s \rangle - 1 + \xi_s \geq 0 \quad (28)$$

$$\forall s \in S: \xi_s \geq 0 \quad (29)$$

$$\forall v \in V: \theta_v^+ \geq 0 \quad (30)$$

$$\forall v \in V: \theta_v^- \geq 0. \quad (31)$$

Summary.

- ▶ Logistic regression is a supervised learning problem with the logistic loss function.
- ▶ The logistic loss can be derived from a probabilistic model with the sigmoid distribution as the likelihood as the normal distribution as the prior.
- ▶ l_2 -regularized linear logistic regression is a convex optimization problem that can be solved, e.g., by the steepest descent algorithm with line search.
- ▶ Support vector machine learning is a supervised learning problem with the hinge loss function.
- ▶ l_1 -regularized linear support vector machine learning is a linear program that can be solved, e.g., by the simplex algorithm (see exercises).