# Machine Learning II

Bjoern Andres
bjoern.andres@tu-dresden.de

Machine Learning for Computer Vision
Faculty of Computer Science
TU Dresden

Version 0.4$\beta$

# Contents

# Chapter 1

# Introduction

## 1.1 Notation

We shall use the following notation:

- We write "iff" as shorthand for "if and only if"
- For any $m \in \mathbb{N}$, we define $[m] = \{0, \dots, m-1\}$.
- For any set $A$, we denote by $2^A$ the power set of $A$
- For any set $A$ and any $m \in \mathbb{N}$, we denote by $\binom{A}{m} = \{B \in 2^A \mid |B| = m\}$ the set of all $m$-elementary subsets of $A$
- For any sets $A, B$, we denote by $B^A$ the set of all maps from $A$ to $B$

# Chapter 2

# Supervised learning

## 2.1 Intuition

Informally, supervised learning is the problem of finding in a family $g : \Theta \to Y^X$ of functions, one $g_\theta : X \to Y$ that minimizes a weighted sum of two objectives:

1. $g$ deviates little from a finite set $\{(x_s, y_s)\}_{s \in S}$ of input-output-pairs
2. $g$ has low complexity, as quantified by a function $R : \Theta \to \mathbb{R}_0^+$

We note that the family $g$ can have meaning beyond a mere parameterization of functions from $X$ to $Y$. For instance, $\Theta$ can be a set of forms, $g$ the functions defined by these forms, and $R$ the length of forms. In that case, supervised learning is really an optimization problem over forms of functions, and $R$ penalizes the complexity of these forms. Moreover, $g$ can be chosen so as to constrain the set of functions from $X$ to $Y$ in the first place.

We concentrate exclusively on the special case where $Y$ is finite. In fact, we concentrate on the case where $Y = \{0, 1\}$ in this chapter and reduce more general cases to this case in Chapter 4.

Moreover, we allow ourselves to take a detour by not optimizing over a family $g : \Theta \to \{0, 1\}^X$ directly but instead optimizing over a family $f : \Theta \to \mathbb{R}^X$ and defining $g$ w.r.t. $f$ via a function $L : \mathbb{R} \times \{0, 1\} \to \mathbb{R}_0^+$, called a *loss function*, such that

$$\forall \theta \in \Theta \; \forall x \in X : \quad g_\theta(x) = \operatorname*{argmin}_{\hat{y} \in \{0,1\}} \; L(f_\theta(x), \hat{y}) \; . \tag{2.1}$$

## 2.2 Definition

**Definition 1** For any $S \neq \varnothing$ finite, called a set of *samples*, any $X \neq \varnothing$, called an *attribute space* and any $x : S \to X$, the tuple $(S, X, x)$ is called *unlabeled data*.

For any $y : S \to \{0, 1\}$, given in addition and called a *labeling*, the tuple $(S, X, x, y)$ is called *labeled data*.

**Definition 2** For any labeled data $T = (S, X, x, y)$, any $\Theta \neq \varnothing$ and family of functions $f : \Theta \to \mathbb{R}^X$, any $R : \Theta \to \mathbb{R}_0^+$, called a *regularizer*, any $L : \mathbb{R} \times \{0, 1\} \to \mathbb{R}_0^+$, called a *loss function*, and any $\lambda \in \mathbb{R}_0^+$, called a *regularization parameter*, the instance of the *supervised learning problem* w.r.t. $T, \Theta, f, R, L$ and $\lambda$ is defined as

$$\inf_{\theta \in \Theta} \quad \lambda R(\theta) + \frac{1}{|S|} \sum_{s \in S} L(f_\theta(x_s), y_s) \tag{2.2}$$

**Definition 3** For any unlabeled data $T = (S, X, x)$, any $\hat{f} : X \to \mathbb{R}$ and any $L : \mathbb{R} \times \{0, 1\} \to \mathbb{R}_0^+$, the instance of the *inference problem* w.r.t. $T, f$ and $L$ is defined as

$$\min_{y' \in \{0,1\}^S} \sum_{s \in S} L(\hat{f}(x_s), y'_s) \tag{2.3}$$

**Lemma 1** *The solutions to the inference problem are the $y : S \to \{0,1\}$ such that*

$$\forall s \in S\colon \quad y_s \in \operatorname*{argmin}_{\hat{y}\in\{0,1\}} L(\hat{f}(x_s), \hat{y}) \ . \tag{2.4}$$

*Moreover, if*

$$\hat{f}(X) \subseteq \{0,1\} \tag{2.5}$$

*and*

$$\forall r \in \mathbb{R}\ \forall \hat{y} \in \{0,1\}\colon \quad L(r,\hat{y}) = \begin{cases} 0 & \text{if } r = \hat{y} \\ 1 & \text{otherwise} \end{cases} \tag{2.6}$$

*then*

$$\forall s \in S\colon \quad y'_s = \hat{f}(x_s) \ . \tag{2.7}$$

Proof Generally, we have

$$\min_{y\in\{0,1\}^S} \sum_{s\in S} L(\hat{f}(x_s), y_s) = \sum_{s\in S} \min_{y_s\in\{0,1\}} L(\hat{f}(x_s), y_s) \tag{2.8}$$

By (2.5), $L(\hat{f}(x_s), \hat{f}(x_s))$ is well-defined for any $s \in S$. By (2.6) and non-negativity of $L$, we have

$$\forall y_s \in \{0,1\}\colon \quad L(\hat{f}(x_s), \hat{f}(x_s)) = 0 \leq L(\hat{f}(x_s), y_s) \ . \tag{2.9}$$

Thus, $y_s = \hat{f}(x_s)$ is optimal for any $s \in S$.

We note that the exact supervised learning problem formalizes a philosophical principle known as Ockham's razor.

# Chapter 3

# Deciding

## 3.1 Linear functions

### 3.1.1 Data

Throughout Section 3.1, we consider real attributes. More specifically, we consider some finite set $V \neq \varnothing$ and labeled data $T = (S, X, x, y)$ with $X = \mathbb{R}^V$. Hence, $x\colon S \to \mathbb{R}^V$ and $y\colon S \to \{0, 1\}$.

### 3.1.2 Familiy of functions

Throughout Section 3.1, we consider linear functions. More specifically, we consider $\Theta = \mathbb{R}^V$ and $f : \Theta \to \mathbb{R}^X$ such that

$$\forall \theta \in \Theta \; \forall \hat{x} \in X\colon \quad f_\theta(\hat{x}) = \langle \theta, \hat{x} \rangle \ . \tag{3.1}$$

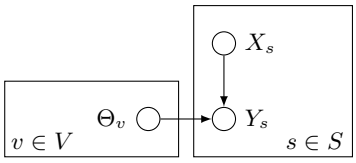
### 3.1.3 Probabilistic model

**Random variables**

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

**Conditional independence assumptions**

We assume a probability distribution that factorizes according to the Bayesian net depicted below.




**Factorization**

- Firstly:

$$P(X, Y, \Theta) = \prod_{s \in S} P(Y_s \mid X_s, \Theta) P(X_s) \prod_{v \in V} P(\Theta_v) \tag{3.2}$$

- Secondly:

$$\begin{aligned} P(\Theta \mid X, Y) &= \frac{P(X, Y, \Theta)}{P(X, Y)} \\ &= \frac{P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(X, Y)} \\ &\propto P(Y \mid X, \Theta)\, P(\Theta) \\ &= \prod_{s \in S} P(Y_s \mid X_s, \Theta) \prod_{v \in V} P(\Theta_v) \end{aligned} \tag{3.3}$$

**Forms**

We consider:

- The *logistic distribution*

$$\forall s \in S: \qquad p_{Y_s|X_s,\Theta}(1) = \frac{1}{1 + 2^{-f_\theta(x_s)}} \tag{3.4}$$

- A $\sigma \in \mathbb{R}^+$ and the *normal distribution*:

$$\forall v \in V: \qquad p_{\Theta_v}(\theta_v) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\theta_v^2/2\sigma^2} \tag{3.5}$$

### 3.1.4 Learning problem

**Lemma 2 (Logistic regression)** *Estimating maximally probable parameters $\theta$, given attributes $x$ and labels $y$, i.e.,*

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

*is identical to the supervised learning problem w.r.t. $L$, $R$ and $\lambda$ such that*

$$\forall r \in \mathbb{R}\ \forall \hat{y} \in \{0, 1\}: \quad L(r, \hat{y}) = -\hat{y}r + \log\left(1 + 2^r\right) \tag{3.6}$$

$$\forall \theta \in \Theta: \quad R(\theta) = \|\theta\|_2^2 \tag{3.7}$$

$$\lambda = \frac{\log e}{2\sigma^2} \tag{3.8}$$

Proof Firstly,

$$\begin{aligned} & \operatorname*{argmax}_{\theta \in \mathbb{R}^m} \quad p_{\Theta|X,Y}(\theta, x, y) \\ \overset{(3.3)}{=} & \operatorname*{argmax}_{\theta \in \mathbb{R}^m} \quad \prod_{s \in S} p_{Y_s|X_s,\Theta}(y_s, x_s, \theta) \prod_{v \in V} p_{\Theta_v}(\theta_v) \\ = & \operatorname*{argmax}_{\theta \in \mathbb{R}^m} \quad \sum_{s \in S} \log p_{Y_s|X_s,\Theta}(y_s, x_s, \theta) + \sum_{v \in V} \log p_{\Theta_v}(\theta_v) \end{aligned} \tag{3.9}$$

Substituting in (3.9) the linearization

$$\begin{aligned} & \log p_{Y_s|X_s,\Theta}(y_s, x_s, \theta) \\ = \quad & y_s \log p_{Y_s|X_s,\Theta}(1, x_s, \theta) + (1 - y_s) \log p_{Y_s|X_s,\Theta}(0, x_s, \theta) \\ = \quad & y_s \log \frac{p_{Y_s|X_s,\Theta}(1, x_s, \theta)}{p_{Y_s|X_s,\Theta}(0, x_s, \theta)} + \log p_{Y_s|X_s,\Theta}(0, x_s, \theta) \end{aligned} \tag{3.10}$$

as well as (3.4) and (3.5) yields the form (3.11) below that is called the instance of the $l_2$-regularized *logistic regression problem* with respect to $x$, $y$ and $\sigma$.

$$\operatorname*{argmin}_{\theta \in \mathbb{R}^m} \quad \sum_{s \in S} \left(-y_s \langle \theta, x_s \rangle + \log\left(1 + 2^{\langle \theta, x_s \rangle}\right)\right) + \frac{\log e}{2\sigma^2} \|\theta\|_2^2 \tag{3.11}$$

**Exercise 1** *a) Derive* (3.11) *from* (3.9) *using* (3.10), (3.4) *and* (3.5)
*b) Is the objective function of* (3.11) *convex?*

### 3.1.5 Inference problem

**Lemma 3** *Estimating maximally probable labels $y$, given attributes $x'$ and parameters $\theta$, i.e.,*

$$\operatorname*{argmax}_{y \in \{0,1\}^S} \quad p_{Y|X,\Theta}(y, x', \theta) \tag{3.12}$$

*is identical to the inference problem w.r.t. $f$ and $L$. It has the solution*

$$\forall s \in S' \; : \quad y_s = \begin{cases} 1 & \text{if } f_\theta(x'_s) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

Proof Firstly,

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

Secondly,

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

### 3.1.6 Inference algorithm

The inference problem is solved by computing independently for each $s \in S'$ the label

$$y_s = \begin{cases} 1 & \text{if } \langle \theta, x'_s \rangle > 0 \\ 0 & \text{otherwise} \end{cases} \; . \tag{3.14}$$

The time complexity is $O(|V||S'|)$.

# Chapter 4

# Semi-supervised and unsupervised learning

## 4.1 Intuition

So far, we have considered learning problems w.r.t. labeled data $(S, X, x, y)$ where, for every $s \in S$, a label $y_s \in \{0, 1\}$ is given, and inference problems w.r.t. unlabeled data $(S', X', x)$ where no label is given and every combination of labels $y' : S \to \{0, 1\}$ is a feasible solution.

Next, we consider learning problems where not every label is given and inference problems where not every combination of labels is feasible. Unlike before, the data we look at in both problems coincides, consisting of tuples $(S, X, x, \mathcal{Y})$ where $\mathcal{Y} \subseteq \{0, 1\}^S$ is a set of feasible labelings. In particular, $\mathcal{Y} = \{0, 1\}^S$ is the special case of unlabeled data, and $|\mathcal{Y}| = 1$ is the special case of labeled data. Non-trivial choices of $\mathcal{Y}$ allow us to express problems of learning and inferring finite structures such as maps (Chapter 5).

## 4.2 Definition

**Definition 4** For any $S \neq \varnothing$ finite, called a set of *samples*, any $X \neq \varnothing$, called an *attribute space*, any $x : S \to X$ and any $\varnothing \neq \mathcal{Y} \subseteq \{0, 1\}^S$, called a set of *feasible labelings*, the tuple $T = (S, X, x, \mathcal{Y})$ is called *constrained data*.

**Definition 5** For any constrained data $T = (S, X, x, \mathcal{Y})$, any $\Theta \neq \varnothing$ and family of functions $f : \Theta \to \mathbb{R}^X$, any $R : \Theta \to \mathbb{R}_0^+$, called a *regularizer*, any $L : \mathbb{R} \times \{0, 1\} \to \mathbb{R}_0^+$, called a *loss function* and any $\lambda \in \mathbb{R}_0^+$, called a *regularization parameter*, the instance of the *learning and inference problem* w.r.t. $T, \Theta, f, R, L$ and $\lambda$ is defined as

$$\min_{y \in \mathcal{Y}} \inf_{\theta \in \Theta} \quad \lambda R(\theta) + \frac{1}{|S|} \sum_{s \in S} L(f_\theta(x_s), y_s) \tag{4.1}$$

The special case of one-elementary $\mathcal{Y} = \{y\}$ is called the *supervised learning problem*.

The special case of one-elementary $\Theta = \{\hat{\theta}\}$ written below is called the *inference problem*.

$$\min_{y \in \mathcal{Y}} \quad \sum_{s \in S} L(f_\theta(x_s), y_s) \tag{4.2}$$

# Chapter 5

# Classifying

## 5.1 Maps

For any finite set $A \neq \varnothing$ whose elements we seek to classify and any finite set $B \neq \varnothing$ of class labels, we are interested in *maps* $\varphi : A \to B$ that assign to every element $a \in A$ precisely one class label $\varphi(a) \in B$. Maps are precisely those subsets of $\varphi \subseteq A \times B$ that satisfy

$$\forall a \in A \; \exists b \in B : \quad (a,b) \in \varphi \tag{5.1}$$

$$\forall a \in A \; \forall b, b' \in B : \quad (a,b) \in \varphi \wedge (a,b') \in \varphi \Rightarrow b = b' \ . \tag{5.2}$$

They are characterized by those functions $y : A \times B \to \{0,1\}$ that satisfy

$$\forall a \in A \colon \quad \sum_{b \in B} y_{ab} = 1 \ . \tag{5.3}$$

We reduce the problem of learning and inferring maps to the problem of learning and inferring decisions, by choosing constrained data with

$$S = A \times B \tag{5.4}$$

$$\mathcal{Y} = \left\{ y \colon A \times B \to \{0,1\} \;\middle|\; \forall a \in A \colon \sum_{b \in B} y_{ab} = 1 \right\} \ . \tag{5.5}$$

## 5.2 Linear functions

### 5.2.1 Data

Throughout Section 5.2, we consider some finite set $V \neq \varnothing$ and constrained data $(S, X, x, \mathcal{Y})$ with $S = A \times B$ as in (5.4), $X = B \times \mathbb{R}^V$, and $\mathcal{Y}$ as in (5.5). More specifically, we assume that, for any $(a,b) \in A \times B$, the class label $b$ is the first attribute of $(a,b)$, i.e.,

$$\forall a \in A \; \forall b \in B \; \exists \hat{x} \in \mathbb{R}^V : \quad x_{ab} = (b, \hat{x}) \tag{5.6}$$

As a special case, we consider labeled data where we are given just one $\mathcal{Y} = \{y\}$ with $y$ satisfying the constraints (5.3).

### 5.2.2 Familiy of functions

Throughout Section 5.2, we consider linear functions. More specifically, we consider $\Theta = \mathbb{R}^{B \times V}$ and $f : \Theta \to \mathbb{R}^X$ such that

$$\forall \theta \in \Theta \; \forall b \in B \; \forall \hat{x} \in \mathbb{R}^V : \quad f_\theta((b, \hat{x})) = \sum_{v \in V} \theta_{bv} \, \hat{x}_v = \langle \theta_{b\cdot}, \hat{x} \rangle \ . \tag{5.7}$$

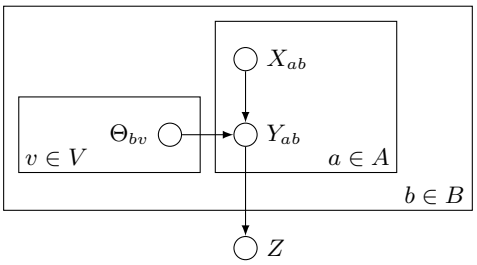
### 5.2.3 Probabilistic model

#### Random variables

- For any $(a, b) \in A \times B$, let $X_{ab}$ be a random variable whose realization is a vector $x_{ab} \in B \times \mathbb{R}^V$, called the *attribute vector* of $(a, b)$.
- For any $(a, b) \in A \times B$, let $Y_{ab}$ be a random variable whose realization is a binary number $y_{ab} \in \{0, 1\}$, called the *decision* of classifying $a$ as $b$
- For any $b \in B$ and any $v \in V$, let $\Theta_{bv}$ be a random variable whose realization is a real number $\theta_{bv} \in \mathbb{R}$, called a *parameter*
- Let $Z$ be a random variable whose realization is a subset $z \subseteq \{0, 1\}^{A \times B}$. For multiple label classification, we are interested in $z = \mathcal{Y}$, the set of the characteristic functions of all maps from $A$ to $B$.

#### Conditional independence assumptions

We assume a probability distribution that factorizes according to Bayesian net depicted below.




#### Factorization

These conditional independence assumptions imply the following factorizations:

- Firstly:

$$P(X, Y, Z, \Theta) = P(Z \mid Y) \prod_{(a,b) \in A \times B} P(Y_{ab} \mid X_{ab}, \Theta) \prod_{(b,v) \in B \times V} P(\Theta_{bv}) \prod_{(a,b) \in A \times B} P(X_{ab}) \tag{5.8}$$

- Secondly:

$$\begin{aligned} P(\Theta \mid X, Y, Z) &= \frac{P(X, Y, Z, \Theta)}{P(X, Y, Z)} \\ &= \frac{P(Z \mid Y)\, P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(Z \mid X, Y)\, P(X, Y)} \\ &= \frac{P(Z \mid Y)\, P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(Z \mid Y)\, P(X, Y)} \\ &= \frac{P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(X, Y)} \\ &\propto P(Y \mid X, \Theta)\, P(\Theta) \\ &= \prod_{(a,b) \in A \times B} P(Y_{ab} \mid X_{ab}, \Theta) \prod_{(b,v) \in B \times V} P(\Theta_{bv}) \end{aligned} \tag{5.9}$$

- Thirdly,

$$\begin{aligned} P(Y \mid X, Z, \theta) &= \frac{P(X, Y, Z, \Theta)}{P(X, Z, \Theta)} \\ &= \frac{P(Z \mid Y)\, P(Y \mid X, \Theta)\, P(X)\, P(\Theta)}{P(X, Z, \Theta)} \\ &\propto P(Z \mid Y)\, P(Y \mid X, \Theta) \\ &= P(Z \mid Y) \prod_{(a,b) \in A \times B} P(Y_{ab} \mid X_{ab}, \Theta) \end{aligned} \tag{5.10}$$

**Forms**

Here, we consider:

- The *logistic distribution*

$$\forall (a,b) \in A \times B: \qquad p_{Y_{ab} | X_{ab}, \Theta}(1) = \frac{1}{1 + 2^{-f_\theta(x_{ab})}} \tag{5.11}$$

- A $\sigma \in \mathbb{R}^+$ and the *normal distribution*:

$$\forall (b,v) \in B \times V: \qquad p_{\Theta_{bv}}(\theta_{bv}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\theta_{bv}^2 / 2\sigma^2} \tag{5.12}$$

- A uniform distribution on a subset:

$$\forall z \subseteq \{0,1\}^{A \times B}: \quad p_{Z|Y}(z) \propto \begin{cases} 1 & \text{if } y \in z \\ 0 & \text{otherwise} \end{cases} \tag{5.13}$$

  Note that $p_{Z|Y}(\mathcal{Y})$ is non-zero iff the relation $y^{-1}(1) \subseteq A \times B$ is a map.

### 5.2.4 Learning problem

**Lemma 4** *Estimating maximally probable parameters $\theta$, given attributes $x$ and decisions $y$, i.e.,*

$$\operatorname*{argmax}_{\theta \in \mathbb{R}^{B \times V}} \quad p_{\Theta | X, Y}(\theta, x, y)$$

*is identical to the supervised learning problem w.r.t. $L$, $R$ and $\lambda$ such that*

$$\forall r \in \mathbb{R}\ \forall \hat{y} \in \{0,1\}: \quad L(r, \hat{y}) = -\hat{y} r + \log\left(1 + 2^r\right) \tag{5.14}$$

$$\forall \theta \in \Theta: \qquad R(\theta) = \|\theta\|_2^2 \tag{5.15}$$

$$\lambda = \frac{\log e}{2\sigma^2} \tag{5.16}$$

*Moreover, this problem separates into $|B|$ independent supervised learning problems, each w.r.t. parameters in $\mathbb{R}^V$, with $L$ and $\lambda$ as above, and with*

$$\forall \theta' \in \mathbb{R}^V: \qquad R'(\theta') = \|\theta'\|_2^2 \tag{5.17}$$

PROOF Analogous to the case of binary classification from Section 3.1, we now obtain:

$$\begin{aligned} &\operatorname*{argmax}_{\theta \in \mathbb{R}^{B \times V}} \quad p_{\Theta | X, Y}(\theta, x, y) \\ &= \operatorname*{argmin}_{\theta \in \mathbb{R}^{B \times V}} \quad \sum_{(a,b) \in A \times B} \left( -y_{ab} f_\theta(x_{ab}) + \log\left(1 + 2^{f_\theta(x_{ab})}\right) \right) + \frac{\log e}{2\sigma^2} \|\theta\|_2^2 \ . \end{aligned} \tag{5.18}$$

Consider the unique $x' : A \times B \to \mathbb{R}^V$ such that, for any $(a, b) \in A \times B$, we have $x_{ab} = (b, x'_{ab})$.

Problem (5.18) separates into $|B|$ many $l_2$-regularized logistic regression problems, one for each $b \in B$, because

$$\begin{aligned}
&\min_{\theta \in \mathbb{R}^{B \times V}} \sum_{(a,b) \in A \times B} \left( -y_{ab} \langle \theta_{b\cdot}, x'_{ab} \rangle + \log \left( 1 + 2^{\langle \theta_{b\cdot}, x'_{ab} \rangle} \right) \right) + \frac{\log e}{2\sigma^2} \|\theta\|_2^2 \\
= &\min_{\theta \in \mathbb{R}^{B \times V}} \sum_{b \in B} \left( \sum_{a \in A} \left( -y_{ab} \langle \theta_{b\cdot}, x'_{ab} \rangle + \log \left( 1 + 2^{\langle \theta_{b\cdot}, x'_{ab} \rangle} \right) \right) + \frac{\log e}{2\sigma^2} \|\theta_{b\cdot}\|_2^2 \right) \\
= &\sum_{b \in B} \min_{\theta_{b\cdot} \in \mathbb{R}^V} \left( \sum_{a \in A} \left( -y_{ab} \langle \theta_{b\cdot}, x'_{ab} \rangle + \log \left( 1 + 2^{\langle \theta_{b\cdot}, x'_{ab} \rangle} \right) \right) + \frac{\log e}{2\sigma^2} \|\theta_{b\cdot}\|_2^2 \right) \quad .
\end{aligned}$$

### 5.2.5 Inference problem

**Lemma 5** *For any constrained data as defined above, any $\theta \in \mathbb{R}^{B \times V}$ and any $\hat{y} : A \times B \to \{0, 1\}$, $\hat{y}$ is a solution to the inference problem*

$$\min_{y \in \mathcal{Y}} \sum_{(a,b) \in A \times B} L(f_\theta(x_{ab}), y_{ab}) \tag{5.19}$$

*iff there exists an $\varphi : A \to B$ such that*

$$\forall a \in A\colon \quad \varphi(a) \in \max_{b \in B} \langle \theta_{b\cdot}, x'_{ab} \rangle \tag{5.20}$$

*and*

$$\forall (a, b) \in A \times B\colon \quad \hat{y}_{ab} = 1 \Leftrightarrow \varphi(a) = b \quad . \tag{5.21}$$

Proof

$$\begin{aligned}
&\sum_{(a,b) \in A \times B} L(f_\theta(x_{ab}), y_{ab}) \\
= &\sum_{(a,b) \in A \times B} \left( L(f_\theta(x_{ab}), 1)\, y_{ab} + L(f_\theta(x_{ab}), 0)\, (1 - y_{ab}) \right) \\
= &\sum_{(a,b) \in A \times B} \left( L(f_\theta(x_{ab}), 1) - L(f_\theta(x_{ab}), 0) \right) y_{ab} + \text{const.} \\
= &\sum_{(a,b) \in A \times B} \left( -f_\theta(x_{ab}) \right) y_{ab} && \text{by (5.14)} \\
= &\sum_{(a,b) \in A \times B} \left( -\langle \theta_{b\cdot}, x'_{ab} \rangle \right) y_{ab} && x_{ab} = (b, x'_{ab}) \\
= &\sum_{a \in A} \sum_{b \in B} \left( -\langle \theta_{b\cdot}, x'_{ab} \rangle \right) y_{ab}
\end{aligned}$$

### 5.2.6 Inference algorithm

The inference problem is solved by solving (5.20) independently for each $a \in A$. The time complexity is $O(|A||B||V|)$.

# Bibliography