Machine Learning 2

A. Advanced Supervised Learning
A.1 Generalized Linear Models

Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL)
Institute for Computer Science
University of Hildesheim, Germany

Syllabus

			A. Advanced Supervised Learning
Fri.	24.4.	(1)	A.1 Generalized Linear Models
Fri.	1.5.	—	— *Labour Day* —
Fri.	8.5.	(2)	A.2 Gaussian Processes
Fri.	15.5.	(3)	A.3 Advanced Support Vector Machines
			B. Ensembles
Fri.	22.5.	(4)	B.1 Stacking
			& B.2 Boosting
Fri.	29.5.	(5)	B.3 Mixtures of Experts
Fri.	5.6.	—	— *Pentecoste Break* —
			C. Sparse Models
Fri.	12.6.	(6)	C.1 Homotopy and Least Angle Regression
Fri.	19.6.	(7)	C.2 Proximal Gradients
Fri.	26.6.	(8)	C.3 Laplace Priors
Fri.	3.7.	(9)	C.4 Automatic Relevance Determination
			D. Complex Predictors
Fri.	10.7.	(10)	D.1 Latent Dirichlet Allocation (LDA)
Fri.	17.7.	(11)	Q & A

Outline

Outline

1. The Prediction Problem / Supervised Learning
2. The Exponential Family
3. Generalized Linear Models (GLMs)
4. Learning Algorithms
5. Organizational Stuff

The Prediction Problem Formally

Let $X_1, X_2, \ldots, X_M$ be random variables called **predictors** (aka **inputs**, **covariates**, **features**), $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_M$ be their domains.

$X := (X_1, X_2, \ldots, X_M)$ the vector of random predictor variables and $\mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_M$ its domain.

Y be a random variable called **target** (or **output**, **response**), $\mathcal{Y}$ be its domain.

$\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$ be a (multi)set of instances of the unknown joint distribution $p(X, Y)$ of predictors and target called **data**. $\mathcal{D}$ is often written as enumeration

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$$

$\mathcal{Y} = \mathbb{R}$: **regression**, $\mathcal{Y}$ a set of nominal values: **classification**.

The Prediction Problem Formally / Test Set Formulation

Let $\mathcal{X}$ be any set (called **predictor space**),
$\mathcal{Y}$ be any set (called **target space**), e.g., and
$p : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_0^+$ be a joint distribution / density.

Given

- a sample $\mathcal{D}^{\text{train}} \subseteq \mathcal{X} \times \mathcal{Y}$ (called **training set**), drawn from p,
- a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that measures how bad it is to predict value $\hat{y}$ if the true value is y,

compute a **model**

$$\hat{y} : \mathcal{X} \to \mathcal{Y}$$

s.t. for another sample $\mathcal{D}^{\text{test}} \subseteq \mathcal{X} \times \mathcal{Y}$ (called **test set**) drawn from the same distribution p, not available during training, the test error

$$\text{err}(\hat{y}; \mathcal{D}^{\text{test}}) := \frac{1}{|D^{\text{test}}|} \sum_{(x,y) \in \mathcal{D}^{\text{test}}} \ell(y, \hat{y}(x))$$

is minimal.

The Prediction Problem Formally / Risk Formulation

Let $\mathcal{X}$ be any set (called **predictor space**),
$\mathcal{Y}$ be any set (called **target space**), and
$p : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_0^+$ be a joint distribution / density.

Given a sample $\mathcal{D}^{\text{train}} \subseteq \mathcal{X} \times \mathcal{Y}$ (called **training set**), drawn from p,
a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ that measures how bad it is to predict value $\hat{y}$ if the true value is y,

compute a **model**

$$\hat{y} : \mathcal{X} \to \mathcal{Y}$$

with minimal risk

$$\text{risk}(\hat{y}; p) := \int_{\mathcal{X} \times \mathcal{Y}} \ell(y, \hat{y})\, p(x, y)\, d(x, y)$$

Explanation: $\text{risk}(\hat{y}; p)$ can be estimated by the **empirical risk**

$$\text{risk}(\hat{y}; \mathcal{D}^{\text{test}}) := \frac{1}{|D^{\text{test}}|} \sum_{(x,y) \in \mathcal{D}^{\text{test}}} \ell(y, \hat{y}(x))$$

Outline

1. The Prediction Problem / Supervised Learning

2. The Exponential Family

3. Generalized Linear Models (GLMs)

4. Learning Algorithms

5. Organizational Stuff

Definition Exponential Family

A parametric pdf $p(\mathrm{x}|\boldsymbol{\theta})$ belongs to the **exponential family** if it is of the form

$$p(\mathrm{x} \mid \boldsymbol{\theta}) = \frac{h(\mathrm{x})}{Z(\boldsymbol{\theta})} e^{\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \Phi(\mathrm{x}) \rangle} = h(\mathrm{x}) e^{\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \Phi(\mathrm{x}) \rangle - A(\boldsymbol{\theta})} \quad (1)$$

- $\boldsymbol{\eta}$ are called **natural** or **canonical** parameters
- $\boldsymbol{\eta}(\boldsymbol{\theta})$ is a **reparametrization**
- $Z(\boldsymbol{\theta}) = \int_{\mathcal{X}} h(\mathrm{x}) e^{\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \Phi(\mathrm{x})} \, \mathrm{dx}$ is called **partition function**
- $A(\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta})$ is called **log partition** or **cumulant** function
- $h(\mathrm{x})$ is a scaling factor called **base measure**
- $\Phi(\mathrm{x})$ is called **sufficient statistic**

Subfamilies

- $\dim(\boldsymbol{\theta}) < \dim \boldsymbol{\eta}(\boldsymbol{\theta})$: **curved exponential family**.
  (more sufficient statistics than parameters)
- $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$: **canonical form**

$$p(x \mid \boldsymbol{\theta}) = h(x) e^{\langle \boldsymbol{\theta}, \Phi(x) \rangle - A(\boldsymbol{\theta})}$$

- $\Phi(x) = x$: **natural exponential family**.

$$p(x \mid \boldsymbol{\theta}) = h(x) e^{\langle \boldsymbol{\eta}(\boldsymbol{\theta}), x \rangle - A(\boldsymbol{\theta})}$$

- natural exponential family in canonical form:

$$p(x \mid \boldsymbol{\theta}) = h(x) e^{\langle \boldsymbol{\theta}, x \rangle - A(\boldsymbol{\theta})}$$

Exponential Distribution?

- **exponential family**:

$$p(\mathrm{x} \mid \boldsymbol{\theta}) = \frac{h(\mathrm{x})}{Z(\boldsymbol{\theta})} e^{\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \Phi(\mathrm{x}) \rangle} = h(\mathrm{x}) e^{\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \Phi(\mathrm{x}) \rangle - A(\boldsymbol{\theta})} \quad (2)$$

- **exponential distribution**:

$$p(x \mid \lambda) := \lambda e^{-\lambda x}$$

- Is the exponential distribution a member of the exponential family?
  A. yes, for $h(x) := e^{-\lambda x}$
  B. yes, for $\eta(\theta) := -\theta$ and $\phi(x) := x$
  C. no, because there is no Z
  B. no, because there is no θ

Examples: Bernoulli

$$\mathcal{X} = \{0, 1\} \qquad \text{Ber}(x \mid \mu) = \mu^x (1-\mu)^{1-x}$$

Examples: Bernoulli

$$\mathcal{X} = \{0,1\} \qquad \text{Ber}(x \mid \mu) = \mu^x (1-\mu)^{1-x}$$

$$\begin{aligned}
&e^{x\log(\mu)+(1-x)\log(1-\mu)}\\
&\theta = \mu\\
&\phi(x) = \begin{pmatrix} x \\ 1-x \end{pmatrix}\\
&\eta(\theta) = \begin{pmatrix} \log\theta \\ \log(1-\theta) \end{pmatrix}\\
\\
&A(\theta) = 0\\
&A(\eta) = 0
\end{aligned} \tag{3}$$

curved

Examples: Bernoulli

$$\mathcal{X} = \{0,1\} \qquad \text{Ber}(x \mid \mu) = \mu^x (1-\mu)^{1-x}$$

$$\begin{array}{ll}
e^{x \log(\mu) + (1-x)\log(1-\mu)} & e^{x \log \frac{\mu}{1-\mu} + \log(1-\mu)} \\
\theta = \mu & \theta = \mu \\
\phi(x) = \begin{pmatrix} x \\ 1-x \end{pmatrix} & \phi(x) = x \\
\eta(\theta) = \begin{pmatrix} \log\theta \\ \log(1-\theta) \end{pmatrix} & \eta(\theta) = \text{logit}(\theta) = \log \frac{\theta}{1-\theta} \\
 & \theta = \text{logistic}(\eta) = \frac{1}{1+e^{-\eta}} \\
A(\theta) = 0 & A(\theta) = -\log(1-\theta) \\
A(\eta) = 0 & A(\eta) = \log(1+e^{\eta}) \\
\textbf{curved} & \textbf{natural}
\end{array} \qquad (3)$$

Examples: Multinoulli / Categorical

$$\mathcal{X} := \{1, 2, \ldots, L\} \equiv \{x \in \{0,1\}^L \mid \sum_{l=1}^{L} x_l = 1\}, \quad \mu \in \Delta_L$$

$$\begin{aligned}
\text{Cat}(x \mid \mu) &:= \prod_{\ell=1}^{L} \mu_\ell^{x_\ell} = e^{\sum_{\ell=1}^{L} x_\ell \log \mu_\ell} \\
&= e^{\sum_{\ell=1}^{L-1} x_\ell \log \mu_\ell + (1 - \sum_{\ell=1}^{L-1} x_\ell)(1 - \sum_{\ell=1}^{L-1} \mu_\ell)} \\
&= e^{\sum_{\ell=1}^{L-1} x_\ell \log \frac{\mu_\ell}{1 - \sum_{\ell'=1}^{L-1} \mu_{\ell'}} + (1 - \sum_{\ell=1}^{L-1} \mu_\ell)} = e^{\eta(\theta)^T x - A(\eta(\theta))}
\end{aligned}$$

$$\phi(x) := x_{1:L-1}, \qquad \theta = \mu_{1:L-1}$$

$$\eta(\theta) := \left(\log \frac{\theta_\ell}{1 - \sum_{\ell'=1}^{L-1} \theta_{\ell'}}\right)_{\ell=1,\ldots,L-1}, \quad \theta(\eta) = \left(\frac{e^{\eta_\ell}}{1 + \sum_{\ell'=1}^{L-1} e^{\eta_{\ell'}}}\right)_{\ell=1}$$

$$A(\eta) := \log(1 + \sum_{\ell=1}^{L-1} e^{\eta_\ell})$$

Note: $\Delta_L := \{\mu \in [0,1]^L \mid \sum_{l=1}^{L} \mu_l = 1\}$ **simplex**, $\text{softmax}(x) := (\frac{e^{x_n}}{\sum_{n=1}^{N} e^{x_n}})_{n=1,\ldots,N}$

Examples: Univariate Gaussian

$$\mathcal{X} := \mathbb{R}$$

$$\mathcal{N}(x \mid \mu, \sigma^2) := \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}} \stackrel{!}{=} e^{\eta(\theta)^T \phi(x) - A(\eta(\theta))}$$

$$\phi(x) := \begin{pmatrix} x \\ x^2 \end{pmatrix}, \quad \theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$$

$$\eta(\theta) := \begin{pmatrix} \theta_1 / \theta_2 \\ -\frac{1}{2\theta_2} \end{pmatrix}$$

$$A(\theta) = \frac{\theta_1^2}{2\theta_2} + \frac{1}{2} \log(2\pi\theta_2)$$

$$\leadsto \quad A(\eta) := -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2} \log(-\eta_2) + \frac{1}{2} \log(\pi), \quad h(x) := 1$$

Non-Examples

Uniform distribution:

$$\text{Unif}(x; a, b) := \frac{1}{b-a}\delta(x \in [a, b])$$

Cumulants

$$\frac{\partial A}{\partial \eta} = E(\phi(x)), \quad \frac{\partial^2 A}{\partial^2 \eta} = \mathrm{var}(\phi(x)), \quad \nabla^2 A(\eta) = \mathrm{cov}(\phi(x))$$

Likelihood and Sufficient Statistics

Data:

$$\mathcal{D} := \{x_1, x_2, \ldots, x_N\}$$

Likelihood:

$$\begin{aligned} p(\mathcal{D} \mid \theta) &= \prod_{n=1}^{N} h(x_n) e^{\eta(\theta)^T \phi(x_n) - A(\eta(\theta))} \\ &= \left(\prod_{n=1}^{N} h(x_n)\right) \left(e^{-A(\eta(\theta))}\right)^N e^{\eta(\theta)^T (\sum_{n=1}^{N} \phi(x_n))} \\ &= \left(\prod_{n=1}^{N} h(x_n)\right) e^{\eta(\theta)^T \phi(\mathcal{D}) - N A(\eta(\theta))}, \quad \phi(\mathcal{D}) := \sum_{n=1}^{N} \phi(x_n) \end{aligned}$$

Maximum Likelihood Estimator (MLE)

$$\log p(\mathcal{D} \mid \theta) = \left(\sum_{n=1}^{N} \log h(x_n)\right) + \eta(\theta)^T \phi(\mathcal{D}) - NA(\eta(\theta))$$

for $h \equiv 1, \eta(\theta) = \theta$:

$$= N + \theta^T \phi(\mathcal{D}) - NA(\theta)$$

$$\frac{\partial \log p}{\partial \theta} = \phi(\mathcal{D}) - N \frac{\partial A(\theta)}{\partial \theta} = \phi(\mathcal{D}) - NE(\phi(x)) \overset{!}{=} 0$$

$$\rightsquigarrow E(\phi(x)) \overset{!}{=} \frac{1}{N} \sum_{n=1}^{N} \phi(x_n) \quad (\textbf{moment matching})$$

Example: Bernoulli

$$\hat{\theta} = \mu := \frac{1}{N} \sum_{n=1}^{N} x_n$$

Why the exponential family matters

- Many common distributions belong to it
- It is the only family of pdfs for which **conjugate priors** exist (later)
- All members of the exponential family are **maximum entropy** pdfs.
- given certain constraints, they are the pdfs. satisfying those constraints which make "the least assumptions about the data"

Outline

Make it Simple

- full exponential family:

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x})\, e^{\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{\phi}(\boldsymbol{x}) - A(\boldsymbol{\theta})}$$

- canonical link ($\eta(\theta) = \theta$), natural sufficient statistics ($\phi(x) = x$):

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x})\, e^{\boldsymbol{x}^T \boldsymbol{\theta} - A(\boldsymbol{\theta})}$$

- and one-dimensional x and θ:

$$p(x \mid \theta) = h(x)\, e^{x\theta - A(\theta)}$$

- and positive $h(x)$:

$$p(x \mid \theta) = e^{x\theta - A(\theta) + c(x)}$$

- But how can we represent a normal distribution this way?

Make it Simple (2/2)

- simplified exponential distribution:

$$p(x \mid \theta) = e^{x\theta - A(\theta) + c(x)}$$

- cannot represent a normal distribution
  - because the sufficient statistics is only one-dimensional, but a normal distribution requires two dimensions
- introduce a parameter again: **dispersion σ^2**:

$$p(x \mid \theta, \sigma^2) = e^{\frac{x\theta - A(\theta)}{\sigma^2} + c(x, \sigma^2)}$$

  - we will see soon, that now a normal distribution can be represented by choosing σ^2 simply as the variance.

Parametrization

- a (simplified) exponential family distribution for the target y:

$$p(y \mid \theta, \sigma^2) := e^{\frac{y\theta - A(\theta)}{\sigma^2} + c(y,\sigma^2)}$$

where σ^2 **dispersion parameter** (often =1),
θ **natural parameter** (a scalar!),
$A(\theta)$ **(log) partition function**,
$c(y, \sigma^2)$ **normalization constant**.

- parametrize θ:

$$\theta = w^T x$$

Parametrization (2/2)

- a (simplified) exponential family distribution for the (one-dimensional regression) target y:

$$p(y \mid x; w, \sigma^2) := e^{\frac{y\, w^T x - A(w^T x)}{\sigma^2} + c(y, \sigma^2)}$$

- subsequently learn w
- but assume σ^2 to be known
  - for normal targets: $\sigma^2 := \text{var}(y)$
  - for most others: $\sigma^2 := 1$

Expectation and Variance

$$\mu = E(y \mid x; w, \sigma^2) = A'(w^T x)$$
$$\tau^2 = \text{Var}(y \mid x; w, \sigma^2) = A''(w^T x)\sigma^2$$

- A' **mean function**, usually denoted by $g^{-1} := A'$
- $\sigma^2 A''$ **variance function**

remember: $p(y \mid x; w, \sigma^2) := e^{\frac{y\, w^T x - A(w^T x)}{\sigma^2} + c(y, \sigma^2)}$

Examples: Linear Regression

$$\mathcal{N}(y; \mu, \sigma^2) := \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad y \in \mathbb{R}$$
$$\mu(x) := w^T x$$

$$\begin{aligned}
\log p(y \mid x, w, \sigma^2) &= -\frac{(y-\mu)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \\
&= \frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)) \\
&= \frac{y\, w^T x - \frac{1}{2}(w^T x)^2}{\sigma^2} - \frac{1}{2}(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)) \\
\rightsquigarrow \quad A(\theta) &= \frac{\theta^2}{2} \\
E(y) &= \mu = w^T x \\
\mathrm{Var}(y) &= \sigma^2
\end{aligned}$$

Examples: Binomial Regression

$$\text{Bin}(y; N, \pi) := \binom{N}{y} \pi^y (1-\pi)^{N-y}, \quad y \in \{0, 1, \ldots, N\}$$
$$\pi(x) := \text{logistic}(w^T x)$$

$$\log p(y \mid x, w) = y \log \frac{\pi}{1-\pi} + N \log(1-\pi) + \log \binom{N}{y}$$
$$\rightsquigarrow \quad A(\theta) = N \log(1 + e^\theta)$$
$$E(y) = \mu = N\pi = N \text{logistic}(w^T x)$$
$$\text{Var}(y) = N\pi(1-\pi) = N \text{logistic}(w^T x)(1 - \text{logistic}(w^T x))$$
$$\text{where } \theta = \log \frac{\pi}{1-\pi} = w^T x$$
$$\sigma^2 = 1$$

Examples: Poisson Regression

$$
\begin{aligned}
\text{Poi}(y;\mu) :=& e^{-\mu}\frac{\mu^y}{y!}, \quad y \in \{0,1,2,\ldots\} \\
\mu(x) :=& e^{w^T x}
\end{aligned}
$$

$$
\begin{aligned}
\log p(y \mid x, w) =& y \log \mu - \mu - \log y! \\
\rightsquigarrow \quad A(\theta) =& e^{\theta} \\
E(y) =& \mu = e^{w^T x} \\
\text{Var}(y) =& e^{w^T x} \\
\text{where } \theta =& \log \mu = w^T x \\
\sigma^2 =& 1
\end{aligned}
$$

Models

Distribution	mean $\mu = g^{-1}(\theta)$	link $\theta = g(\mu)$
$\mathcal{N}(y; \mu, \sigma^2)$	$\mu = g^{-1}(\theta) = \theta$	$\theta = g(\mu) = \mu$
$\text{Bin}(y; N, \mu)$	$\mu = g^{-1}(\theta) = N \,\text{logistic}(\theta)$	$\theta = g(\mu) = \text{logit}(\frac{\mu}{N})$
$\text{Poi}(y; \mu)$	$\mu = g^{-1}(\theta) = e^{\theta}$	$\theta = g(\mu) = \log \mu$
$\text{Ber}(y; \mu)$	$\mu = g^{-1}(\theta) = \text{logistic}(\theta)$	$\theta = g(\mu) = \text{logit}(\mu)$

Outline

Gradient Descent

model:

$$p(y \mid x; w, \sigma^2) := e^{\frac{y\, w^T x - A(w^T x)}{\sigma^2} + c(y, \sigma^2)}$$

$$\text{with } \theta = w^T x$$

negative log likelihood:

$$\ell(w; x, y) = -\sum_{n=1}^{N} \frac{y_n\, w^T x_n - A(w^T x_n)}{\sigma^2} =: -\frac{1}{\sigma^2} \sum_{n=1}^{N} \ell_n(w^T x_n)$$

$$\frac{\partial \ell_n}{\partial w_m} = \frac{\partial \ell_n}{\partial \theta_n} \frac{\partial \theta_n}{\partial w_m} = (y_n - \mu_n) \frac{\partial \theta_n}{\partial w_m} = (y_n - \mu_n) x_{n,m}$$

and thus:

$$\nabla_w \ell(w) = -\frac{1}{\sigma^2} \sum_{n=1}^{N} (y_n - \mu_n) x_n$$

Newton

$$\nabla_w \ell(w) = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(y_n - \mu_n)x_n$$

$$\frac{\partial^2 \ell}{\partial^2 w} = \frac{1}{\sigma^2}\sum_{n=1}^{N}\frac{\partial \mu_n}{\partial \theta_n} x_n x_n^T = \frac{1}{\sigma^2} X^T S X$$

$$\text{where } S := \text{diag}\left(\frac{\partial \mu_1}{\partial \theta_1}, \ldots, \frac{\partial \mu_N}{\partial \theta_N}\right)$$

Use within IRLS:

$$\theta^{(t)} := X w^{(t)}$$

$$\mu^{(t)} := g^{-1}(\theta^{(t)})$$

$$z^{(t)} := \theta^{(t)} + (S^{(t)})^{-1}(y - \mu^{(t)})$$

$$w^{(t+1)} := (X^T S^{(t)} X)^{-1} X^T S^{(t)} z^{(t)}$$

Stochastic Gradient Descent

$$\nabla_w \ell(w) = -\frac{1}{\sigma^2} \sum_{n=1}^{N} (y_n - \mu_n) x_n$$

Use a smaller subset of data to estimate the (stochastic) gradient:

$$\nabla_w \ell(w) \approx -\frac{1}{\sigma^2} \sum_{n \in S} (y_n - \mu_n) x_n, \quad S \subseteq \{1, \ldots, N\}$$

Extreme case: use only one sample at a time (online):

$$\nabla_w \ell(w) \approx -\frac{1}{\sigma^2} (y_n - \mu_n) x_n, \quad n \in \{1, \ldots, N\}$$

Beware: $\nabla_w \ell(w) \approx 0$ then is not a useful stopping criterion!

L2 Regularization

For all models, do not forget to add L2 regularization.

Straight-forward to add to all learning algorithms discussed.

Outline

Character of the Lecture

This is an advanced lecture:

- I will assume good knowledge of Machine Learning I.
- Slides will contain major keywords, not the full story.
- For the full story, you need to read the referenced chapters in one of the books.

Exercises and Tutorials

- Two tutorials (course 3102):
  - Tuesdays, 8:00-10:00 - Zoom 920-3439-1690
  - Wednesdays, 14:00-16:00 - Zoom 961-3756-7779
- Tutorial sheet upload: each **Friday, 12:00 AM** (LearnWeb)
- Tutorial sheet deadline: **Thursday, 12:00 AM**
  (upload in Learnweb as PDF)
- 50% Tutorial sheet points required to PASS the course
  No exam or final grade this year
- Next week: Tutorial via Zoom
  Check Learnweb or our ISMLL-Website for the Meeting-ID.

Passing Requirements

- 50% Tutorial sheet points
- Participation in class:
  Present your Solution at least twice
- No Group Submissions allowed
- **Plagiarism** will lead to immediately failing the course

Credit Points

- The course gives 6 ECTS (2+2 SWS).
- The course can be used in
  - IMIT MSc. / Informatik / Gebiet KI & ML
  - Wirtschaftsinformatik MSc / Informatik / Gebiet KI & ML & Wirtschaftsinformatik MSc / Wirtschaftsinformatik / Gebiet BI
  - as well as in both BSc programs.

Some Books

- Kevin P. Murphy (2012):
  Machine Learning, A Probabilistic Approach, MIT Press.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman (^{2}2009):
  The Elements of Statistical Learning, Springer.

  Also available online as PDF at `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`
- Christopher M. Bishop (2007):
  Pattern Recognition and Machine Learning, Springer.
- Richard O. Duda, Peter E. Hart, David G. Stork (^{2}2001):
  Pattern Classification, Springer.

Summary

- Generalized linear models allow to model regression targets with
  - specific domains: $\mathbb{R}$, $\mathbb{R}_0^+$, $\{0, 1\}$, $\{1, \ldots, K\}$, $\mathbb{N}_0$ etc.
  - specific parametrized shapes of pdfs/pmfs.
- The model is composed of
  1. a linear combination of the predictors and
  2. a scalar transform to the domain of the target (**mean function**, inverse **link function**)
- Many well-known models are special cases of GLMs:
  - linear regression (= GLM with normally distributed target)
  - logistic regression (= GLM with bernoulli distributed target)
  - Poisson regression (= GLM with Poisson distributed target)
- Generic simple learning algorithms exist for GLMs independent of the target distribution.
- GLMs have a principled probabilistic interpretation and provide posterior distributions (uncertainty/risk).

Further Readings

- See also [**?**, chapter 9].

References

Kevin P. Murphy.
Machine learning: a probabilistic perspective.
The MIT Press, 2012.