Machine Learning 2

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
Tue. 5.4.	(1)	A.1 Generalized Linear Models
Tue. 12.4.	(2)	A.2 Gaussian Processes
Tue. 19.4.	(3)	A.3 Advanced Support Vector Machines
Tue. 26.4.	(4)	A.4 Neural Networks
Tue. 3.5.	(5)	A.5 Ensembles
Tue. 10.5.	(6)	A.5b Ensembles (ctd.)
Tue. 17.5.	(7)	A.6 Sparse Linear Models — L1 regularization
Tue. 24.5.	—	— Pentecoste Break —
Tue. 31.5.	(8)	A.6b Sparse Linear Models — L1 regularization (ctd.)
Tue. 7.6.	(9)	A.7. Sparse Linear Models — Further Methods
		B. Complex Predictors
Tue. 14.6.	(10)	B.1 Latent Dirichlet Allocation (LDA)
Tue. 21.6.	(11)	B.1b Latent Dirichlet Allocation (LDA; ctd.)
Tue. 28.6.	(12)	B.2 Deep Learning
Tue. 5.7.	(13)	Questions and Answers

Outline

1. The Exponential Family

2. Generalized Linear Models (GLMs)

3. Learning Algorithms

Outline

1. The Exponential Family
2. Generalized Linear Models (GLMs)
3. Learning Algorithms

Definition Exponential Family

Let $\mathcal{X}$ be a set,

$\phi : \mathcal{X} \rightarrow \mathbb{R}^M$ a function called **sufficient statistics**,

$h : \mathcal{X} \rightarrow \mathbb{R}$ a function called **scaling function**, often $h \equiv 1$,

$\eta : \mathbb{R}^K \rightarrow \mathbb{R}^M$ a function called **natural parameter**,

then the pdf / pmf

$$p(x \mid \theta) := \frac{1}{Z(\eta(\theta))} h(x) e^{\eta(\theta)^T \phi(x)}$$

with $Z(\theta) := \int_{\mathcal{X}} h(x) e^{\theta^T \phi(x)} dx$ called **partition function**

is called **a member of the exponential family**.

$\theta \in \mathbb{R}^M$ are called **parameters**.

Definition Exponential Family

Let $\mathcal{X}$ be a set,

$\phi : \mathcal{X} \to \mathbb{R}^M$ a function called **sufficient statistics**,

$h : \mathcal{X} \to \mathbb{R}$ a function called **scaling function**, often $h \equiv 1$,

$\eta : \mathbb{R}^K \to \mathbb{R}^M$ a function called **natural parameter**,

then the pdf / pmf

$$
\begin{aligned}
p(x \mid \theta) :=& \frac{1}{Z(\eta(\theta))} h(x) e^{\eta(\theta)^T \phi(x)} \\
=& h(x) e^{\eta(\theta)^T \phi(x) - A(\eta(\theta))}
\end{aligned}
$$

with $Z(\theta) := \int_{\mathcal{X}} h(x) e^{\theta^T \phi(x)} dx$ called **partition function**

$A(\theta) := \log Z(\theta)$ called **log partition function / cumulant**

is called **a member of the exponential family**.

$\theta \in \mathbb{R}^M$ are called **parameters**.

Subfamilies

$K < M$: **curved exponential family**.

$\eta(\theta) = \theta$: **canonical form**:

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

$\phi(x) = x, \mathcal{X} = \mathbb{R}^M$: **natural exponential family**:

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

natural exponential family in canonical form:

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

Examples: Bernoulli

$$
\begin{aligned}
\mathcal{X} :=& \{0,1\} \\
\mathrm{Ber}(x \mid \mu) :=& \mu^x (1-\mu)^{1-x} \\
=& e^{x \log(\mu) + (1-x)\log(1-\mu)} \\
=& e^{\eta(\theta)^T \phi(x)}, \\
\phi(x) :=& \begin{pmatrix} x \\ 1-x \end{pmatrix}, \\
\eta(\theta) :=& \begin{pmatrix} \log\theta \\ \log(1-\theta) \end{pmatrix} \\
A(\eta) :=& 0 \\
\theta =& \mu
\end{aligned}
$$

Examples: Bernoulli

$$
\begin{aligned}
\mathcal{X} :=& \{0,1\} \\
\mathrm{Ber}(x \mid \mu) :=& \mu^x (1-\mu)^{1-x} \\
=& e^{x \log(\mu) + (1-x)\log(1-\mu)} \\
=& e^{\eta(\theta)^T \phi(x)}, \\
\phi(x) :=& \begin{pmatrix} x \\ 1-x \end{pmatrix}, \\
\eta(\theta) :=& \begin{pmatrix} \log\theta \\ \log(1-\theta) \end{pmatrix} \\
A(\eta) :=& 0 \\
\theta =& \mu
\end{aligned}
$$

Linear dependency in $\phi(x)$: $\begin{pmatrix} 1 \\ 1 \end{pmatrix}^T \phi(x) = 1$ (**over-complete**)

Examples: Bernoulli

$$\begin{aligned}
\mathcal{X} :=& \{0,1\} \\
\mathrm{Ber}(x \mid \mu) :=& \mu^x (1-\mu)^{1-x} \\
=& e^{x \log(\mu) + (1-x)\log(1-\mu)} = e^{x \log \frac{\mu}{1-\mu} + \log(1-\mu)} \\
=& e^{\eta(\theta)^T x - A(\eta(\theta))}, \\
\phi(x) :=& x, \\
\eta(\theta) :=& \log \frac{\theta}{1-\theta}, \quad \theta = \mathrm{logistic}(\eta) := \frac{e^\eta}{1+e^\eta} \\
A(\eta) :=& \log(1+e^\eta) \\
\theta =& \mu
\end{aligned}$$

Examples: Multinoulli

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

$$\begin{aligned}
\mathrm{Cat}(x \mid \mu) :=& \prod_{\ell=1}^{L} \mu_\ell^{x_\ell} = e^{\sum_{\ell=1}^{L} x_\ell \log \mu_\ell} \\
=& e^{\sum_{\ell=1}^{L-1} x_\ell \log \mu_\ell + (1 - \sum_{\ell=1}^{L-1} x_\ell)(1 - \sum_{\ell=1}^{L-1} \mu_\ell)} \\
=& e^{\sum_{\ell=1}^{L-1} x_\ell \log \frac{\mu_\ell}{1 - \sum_{\ell'=1}^{L-1} \mu_{\ell'}} + (1 - \sum_{\ell=1}^{L-1} \mu_\ell)} = e^{\eta(\theta)^T x - A(\eta(\theta))} \\
\phi(x) :=& x_{1:L-1} \\
\eta(\theta) :=& \left(\log \frac{\theta_\ell}{1 - \sum_{\ell'=1}^{L-1} \theta_{\ell'}} \right)_{\ell = 1, \ldots, L-1} \\
A(\eta) :=& \log(1 + \sum_{\ell=1}^{L-1} e^{\eta_\ell}), \quad \theta = \mu_{1:L-1}
\end{aligned}$$

Examples: Univariate Gaussian

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

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

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

$$\phi(x) := \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

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

$$A(\eta) := -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2) - \frac{1}{2}\log(2\pi), \quad \theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$$

$$h(x) := \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}$$

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)

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

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

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

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

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

Example: Bernoulli

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

Outline

Parametrization

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

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

Model

$$
\begin{tikzcd}
\boldsymbol{w} \arrow[dr] & & & \\
& \eta_i \arrow[r, shift left, "g^{-1}"] & \mu_i \arrow[l, shift left, "g"] \arrow[r, shift left, "\Psi"] & \theta_i \arrow[l, shift left, "\Psi^{-1}"] \\
\boldsymbol{x}_i \arrow[ur] & & &
\end{tikzcd}
$$

Model with canonical link ($g = \psi$)

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

setting

$$\theta = w^T x$$

Models

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

Expectation and Variance

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

Examples: Linear Regression

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

$$\mu(x) := w^T x$$

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

$$\begin{aligned}
\rightsquigarrow \quad A(\theta) &= \frac{\theta^2}{2} \\
E(y) &= \mu = w^T x \\
\mathrm{Var}(y) &= \sigma^2
\end{aligned}$$

Examples: Binomial Regression

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

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

Examples: Poisson Regression

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

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

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:

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

$$\frac{\partial \ell_n}{\partial w_m} = \frac{\partial \ell_n}{\partial \theta_n} \frac{\partial \theta_n}{\partial \mu_n} \frac{\partial \mu_n}{\partial \eta_n} \frac{\partial \eta_n}{\partial w_m}$$

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

and thus with canonical link:

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

Newton

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

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

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

Use within IRLS:

$$\begin{aligned}
\theta^{(t)} &:= X w^{(t)} \\
\mu^{(t)} &:= g^{-1}(\theta^{(t)}) \\
z^{(t)} &:= \theta^{(t)} + (S^{(t)})^{-1}(y - \mu^{(t)}) \\
w^{(t+1)} &:= (X^T S^{(t)} X)^{-1} X^T S^{(t)} z^{(t)}
\end{aligned}$$

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.

Summary

- Generalized linear models allow to model 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 binomially 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 [Mur12, chapter 9].

References

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