The Exponential Family & Generalized Linear Models

CSC2541 Tutorial 2, Winter 2022

Jenny Bao

Jan 20, 2022

Overview

The Exponential Family

- Formula & basics
- Examples: Bernoulli, Gaussian, ...
- Useful identities

Generalized Linear Models

The Exponential Family

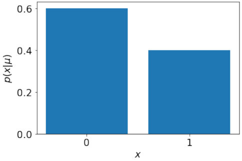
$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}))$$

- $\boldsymbol{\eta}$: natural parameters
- $\boldsymbol{u}(\boldsymbol{x})$: sufficient statistic
- $\mathcal{Z}(\boldsymbol{\eta})$: partition function, ensures the distribution $p(\boldsymbol{x}|\boldsymbol{\eta})$ is normalized.

$$\text{continuous:} \quad \mathcal{Z}(\boldsymbol{\eta}) = \int h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x})) d\boldsymbol{x}$$

$$\text{discrete:} \quad \mathcal{Z}(\boldsymbol{\eta}) = \sum_{\boldsymbol{x}} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}))$$

Example 1: Bernoulli



$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}))$$

Bernoulli distribution:

$$p(x|\mu) = \mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$$

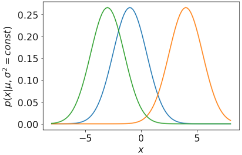
Put in the exponential family form:

$$\begin{aligned}\mathrm{Bern}(x|\mu) &= (1-\mu)\left(\frac{\mu}{1-\mu}\right)^x \\ &= \underbrace{(1-\mu)}_{\frac{1}{\mathcal{Z}(\eta)}} \cdot \underbrace{1}_{h(x)} \cdot \exp\left\{ \Big(\underbrace{\log \frac{\mu}{1-\mu}}_{\eta} \Big) \cdot \underbrace{x}_{\boldsymbol{u}(x)} \right\}\end{aligned}$$

- $\implies \mu = \sigma(\eta),\ \mathcal{Z}(\eta) = \sigma(-\eta)$

Exercise: put the multinomial distribution in the standard form for exponential family

Example 2: Gaussian (μ)

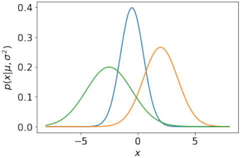


$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}))$$

Gaussian distribution (treating only μ as parameter, assuming σ is constant):

$$\begin{aligned}
p(x|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}) \\
&= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{x^2}{2\sigma^2})}_{h(\boldsymbol{x})} \cdot \underbrace{\exp(-\frac{\mu^2}{2\sigma^2})}_{\frac{1}{\mathcal{Z}(\boldsymbol{\eta})}} \exp(\underbrace{\frac{\mu}{\sigma^2}}_{\boldsymbol{\eta}^\top} \underbrace{x}_{\boldsymbol{u}(\boldsymbol{x})})
\end{aligned}$$

Example 3: Gaussian (μ and σ)



$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}))$$

Gaussian distribution (treating both μ and σ as parameters):

$$\begin{aligned}
p(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}} \sigma^{-1} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2) \\
&= \frac{1}{\sqrt{2\pi}} \sigma^{-1} \exp(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}) \\
&= \underbrace{\frac{1}{\sqrt{2\pi}}}_{h(\boldsymbol{x})} \cdot \underbrace{\sigma^{-1}\exp(-\frac{\mu^2}{2\sigma^2})}_{\frac{1}{\mathcal{Z}(\boldsymbol{\eta})}} \exp(\underbrace{\begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}^\top}_{\boldsymbol{\eta}^\top} \underbrace{\begin{bmatrix} x \\ x^2 \end{bmatrix}}_{\boldsymbol{u}(\boldsymbol{x})})
\end{aligned}$$

The Exponential Family

Other members of the exponential family:

- Poisson, gamma, exponential, beta, Dirichlet, ...

Why studying the exponential family?

Many convenient properties

- Sufficient statistics for maximum likelihood
- Many convenient identities for $\mathcal{Z}(\boldsymbol{\eta})$ (the partition function)
  - Relates concepts such as the Fisher information matrix

Can be used to derive the Generalized Linear Models (GLM)

Maximum likelihood & sufficient statistics

Consider i.i.d. data $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. Find $\boldsymbol{\eta}$ to maximize $p(\boldsymbol{X}|\boldsymbol{\eta})$ (maximum likelihood).

$$
\begin{aligned}
p(\boldsymbol{X}|\boldsymbol{\eta}) &= \prod_{i=1}^{N} \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} \left(h(\boldsymbol{x}_i) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}_i)) \right) \\
&= \left(\frac{1}{\mathcal{Z}(\boldsymbol{\eta})} \right)^N \left(\prod_{i=1}^{N} h(\boldsymbol{x}_i) \right) \exp(\boldsymbol{\eta}^\top \sum_{i=1}^{N} \boldsymbol{u}(\boldsymbol{x}_i))
\end{aligned}
$$

Take derivative of the log-likelihood and set it to 0.

$$
\begin{aligned}
\nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{X}|\boldsymbol{\eta}) = -N \nabla_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) + \sum_{i=1}^{N} \boldsymbol{u}(\boldsymbol{x}_i) = 0 \\
\Longrightarrow \nabla_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}(\boldsymbol{x}_i)
\end{aligned}
$$

Maximum likelihood & sufficient statistics

$$\nabla_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}(\boldsymbol{x}_i)$$

The maximum-likelihood solution $\boldsymbol{\eta}$ only depends on $\sum_{i=1}^{N} \boldsymbol{u}(\boldsymbol{x}_i)$.

- Hence $\boldsymbol{u}(\boldsymbol{x})$ is called the sufficient statistic

Examples:

- **Bernoulli:** $u(x) = x$. Only need to store $\sum_i x_i$.
- **Gaussian (μ and σ):** $u(x) = \begin{bmatrix} x & x^2 \end{bmatrix}^\top$. Need to store both $\sum_i x_i$ and $\sum_i x_i^2$.

Identities

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}))$$

1. $\boxed{\nabla_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) = \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{\eta})}[\boldsymbol{u}(\boldsymbol{x})] =: \boldsymbol{\xi}}$ (moments)

Derivation:

$$\begin{aligned}
\nabla_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) &= \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} \nabla_{\boldsymbol{\eta}} \int h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x})) d\boldsymbol{x} \\
&= \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} \int h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x})) \boldsymbol{u}(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x})) \boldsymbol{u}(\boldsymbol{x}) d\boldsymbol{x} \\
&= \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{\eta})}[\boldsymbol{u}(\boldsymbol{x})]
\end{aligned}$$

Identities

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}))$$

1. $\boxed{\nabla_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) = \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{\eta})}[\boldsymbol{u}(\boldsymbol{x})] =: \boldsymbol{\xi}}$ (moments)

- There's a 1-to-1 mapping between $\boldsymbol{\eta} \leftrightarrow \boldsymbol{\xi}$
- $\boldsymbol{\xi}$ is an alternative parameterization for the exponential family

$$p(\boldsymbol{x}|\boldsymbol{\eta}) \leftrightarrow p(\boldsymbol{x}|\boldsymbol{\xi})$$

Identities

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}))$$

1. $\boxed{\nabla_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) = \mathbb{E}_{x \sim p(x|\boldsymbol{\eta})}[\boldsymbol{u}(\boldsymbol{x})] =: \boldsymbol{\xi}}$ (moments)
2. $\boxed{\nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{x}|\boldsymbol{\eta}) = \boldsymbol{u}(\boldsymbol{x}) - \mathbb{E}_{x \sim p(x|\boldsymbol{\eta})}[\boldsymbol{u}(\boldsymbol{x})]}$

Derivation:

$$\begin{aligned}
\nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{x}|\boldsymbol{\eta}) &= -\nabla_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) + \nabla_{\boldsymbol{\eta}}(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x})) \\
&= -\mathbb{E}_{x \sim p(x|\boldsymbol{\eta})}[\boldsymbol{u}(\boldsymbol{x})] + \boldsymbol{u}(\boldsymbol{x})
\end{aligned}$$

Identities

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}))$$

1. $\boxed{\nabla_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) = \mathbb{E}_{x\sim p(x|\boldsymbol{\eta})}[\boldsymbol{u}(\boldsymbol{x})] =: \boldsymbol{\xi}}$ (moments)
2. $\boxed{\nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{x}|\boldsymbol{\eta}) = \boldsymbol{u}(\boldsymbol{x}) - \mathbb{E}_{x\sim p(x|\boldsymbol{\eta})}[\boldsymbol{u}(\boldsymbol{x})]}$

Recall, in maximum likelihood, we have

$$\nabla_{\boldsymbol{\eta}} \sum_{i=1}^{N} \log p(\boldsymbol{x}_i|\boldsymbol{\eta}) = N\left(\underbrace{\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{u}(\boldsymbol{x}_i)}_{\text{empirical moments } \hat{\boldsymbol{\xi}}} - \underbrace{\nabla_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta})}_{\text{moments } \boldsymbol{\xi}} \right)$$

$$= N \cdot (\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}) = 0$$

Maximum likelihood $\rightarrow$ moment matching

Identities

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}))$$

1. $\boxed{\nabla_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) = \mathbb{E}_{x \sim p(x|\eta)}[\boldsymbol{u}(\boldsymbol{x})] =: \boldsymbol{\xi}}$ (moments)
2. $\boxed{\nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{x}|\boldsymbol{\eta}) = \boldsymbol{u}(\boldsymbol{x}) - \mathbb{E}_{x \sim p(x|\eta)}[\boldsymbol{u}(\boldsymbol{x})]}$

3.1. $\boxed{\nabla^2_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) = \mathrm{Cov}(\boldsymbol{u}(\boldsymbol{x})) = -\nabla^2_{\boldsymbol{\eta}} \log p(\boldsymbol{x}|\boldsymbol{\eta})}$

Derivation: $\nabla^2_{\boldsymbol{\eta}} \log p(\boldsymbol{x}|\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}} \boldsymbol{u}(\boldsymbol{x}) - \nabla^2_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) = -\nabla^2_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta})$

$$\begin{aligned}
\nabla^2_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) &= \nabla_{\boldsymbol{\eta}} \mathbb{E}_{x \sim p(x|\eta)}[\boldsymbol{u}(\boldsymbol{x})] \\
\text{the grad-log trick} \rightarrow \quad &= \mathbb{E}_{x \sim p(x|\eta)}[\nabla_\eta \log p(\boldsymbol{x}|\boldsymbol{\eta}) \boldsymbol{u}(\boldsymbol{x})^\top] \\
&= \mathbb{E}_{x \sim p(x|\eta)}[(\boldsymbol{u}(\boldsymbol{x}) - \mathbb{E}_{x \sim p(x|\eta)}[\boldsymbol{u}(\boldsymbol{x})]) \boldsymbol{u}(\boldsymbol{x})^\top] \\
\mathbb{E}[(\boldsymbol{u} - \mathbb{E}[\boldsymbol{u}])\mathbb{E}[\boldsymbol{u}]] = 0 \rightarrow \quad &= \mathbb{E}[(\boldsymbol{u}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{u}(\boldsymbol{x})])(\boldsymbol{u}(\boldsymbol{x}) - \mathbb{E}[\boldsymbol{u}(\boldsymbol{x})])^\top] \\
&= \mathrm{Cov}(\boldsymbol{u}(\boldsymbol{x}))
\end{aligned}$$

Identities

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}))$$

1. $\boxed{\nabla_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) = \mathbb{E}_{x \sim p(x|\boldsymbol{\eta})}[\boldsymbol{u}(\boldsymbol{x})] =: \boldsymbol{\xi}}$ (moments)
2. $\boxed{\nabla_{\boldsymbol{\eta}} \log p(\boldsymbol{x}|\boldsymbol{\eta}) = \boldsymbol{u}(\boldsymbol{x}) - \mathbb{E}_{x \sim p(x|\boldsymbol{\eta})}[\boldsymbol{u}(\boldsymbol{x})]}$

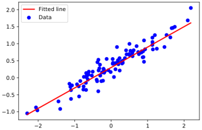
3.1. $\boxed{\nabla^2_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) = \mathrm{Cov}(\boldsymbol{u}(\boldsymbol{x})) = -\nabla^2_{\boldsymbol{\eta}} \log p(\boldsymbol{x}|\boldsymbol{\eta})}$

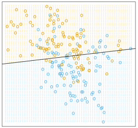
3.2. $\boxed{\nabla^2_{\boldsymbol{\eta}} \log \mathcal{Z}(\boldsymbol{\eta}) = F_{\boldsymbol{\eta}}}$ (Fisher information matrix)

3.3. $\boxed{\boldsymbol{F}_{\boldsymbol{\eta}} = \nabla_{\boldsymbol{\eta}} \boldsymbol{\xi} = \boldsymbol{J}_{\boldsymbol{\xi},\boldsymbol{\eta}}}$ (Jacobian of mapping $\boldsymbol{\eta} \to \boldsymbol{\xi}$)

We will skip the details for now, as the Fisher information matrix will be covered in lecture 3. Also, discussion on these identities are in Chapter 3 of the course notes.

Generalized Linear Models





Consider two familiar models

Linear regression: $\mathcal{L} = \frac{1}{2N}\sum_{i=1}^{N}(t_i - y_i)^2$, where $y_i = \boldsymbol{w}^\top \boldsymbol{x}_i$

$$\implies \nabla_{\boldsymbol{w}}\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}(t_i - y_i)x_i$$

Logistic regression: $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} -t_i \log y_i - (1 - t_i)\log(1 - y_i)$, where $y_i = \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i)$

$$\implies \nabla_{\boldsymbol{w}}\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}(t_i - y_i)x_i$$

Their gradients have the same form (!!!) Why?

(Images taken from CSC311 lecture slides).

Generalized Linear Models

- Linear regression, logistic regression, softmax regression all belong to a broader class of models called **generalized linear models (GLM)**.
- GLM is derived from the exponential family.

Generalized Linear Models

Consider the linear model with features:

$$
\begin{aligned}
z &= \boldsymbol{w}^\top \phi(\boldsymbol{x}) \\
y &= a(z) \qquad\qquad \text{(activation)}
\end{aligned}
$$

Assume the labels are distributed according to the exponential family (implied in the loss function)

$$p(t|\eta) = \frac{1}{\mathcal{Z}(\eta)} h(t) \exp(\eta^\top u(t))$$

We focus on a special case of the exponential family where $u(t) = t$.

$$p(t|\eta) = \frac{1}{\mathcal{Z}(\eta)} h(t) \exp(\eta^\top t)$$

Generalized Linear Models

$$z = \boldsymbol{w}^\top \phi(\boldsymbol{x}), \quad y = a(z)$$

$$p(t|\eta) = \frac{1}{\mathcal{Z}(\eta)} h(t) \exp(\eta^\top t)$$

Recall, the moments can be computed by differentiating the partition function:

$$\nabla_\eta \log \mathcal{Z}(\eta) = \mathbb{E}_{t \sim p(t|\eta)}[u(t)] = \mathbb{E}_{t \sim p(t|\eta)}[t] = y \qquad \text{(Prediction probability)}$$

There is a 1-to-1 mapping between $\eta \leftrightarrow y$. Let

$$\eta = \psi(y)$$

Generalized Linear Models

$$z = \boldsymbol{w}^\top \phi(\boldsymbol{x}), \quad y = a(z)$$
$$\eta = \psi(y)$$
$$p(t|\eta) = \frac{1}{\mathcal{Z}(\eta)} h(t) \exp(\eta^\top t)$$

Gradient of log-likelihood w.r.t weights $\boldsymbol{w}$:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{w}} \sum_{i=1}^{N} \log p(t_i|\eta_i) &= \sum_{i=1}^{N} \frac{\partial}{\partial \eta_i} \log p(t_i|\eta_i) \frac{\partial \eta_i}{\partial y_i} \frac{\partial y_i}{\partial z_i} \frac{\partial z_i}{\partial \boldsymbol{w}} \\
&= \sum_{i=1}^{N} (u(t_i) - \mathbb{E}_{t_i \sim p(t_i|\eta_i)}[u(t_i)]) \psi'(y_i) a'(z_i) \phi(\boldsymbol{x_i}) \\
&= \sum_{i=1}^{N} (t_i - y_i) \psi'(y_i) a'(z_i) \phi(\boldsymbol{x_i})
\end{aligned}$$

This greatly simplifies if we choose $a = \psi^{-1}$.

$$\eta = \psi(a(z)) = \psi(\psi^{-1}(z)) = z \implies \psi'(y) a'(z) = \frac{\partial \eta}{\partial y} \frac{\partial y}{\partial z} = 1$$

Generalized Linear Models

To summarize, when the following conditions are met:

- Linear model with activation

$$z = \boldsymbol{w}^\top \phi(\boldsymbol{x}), \quad y = a(z)$$

- The label distribution belongs to the exponential family

$$p(\boldsymbol{t}|\eta) = \frac{1}{\mathcal{Z}(\eta)} h(\boldsymbol{t}) \exp(\eta^\top \boldsymbol{t})$$
$$\text{where } y = \nabla_\eta \log \mathcal{Z}(\eta), \quad \text{and } \eta = \psi(y)$$

- Activation is chosen as:

$$a(\cdot) = \psi^{-1}(\cdot)$$

Then we have:

$$\frac{\partial}{\partial \boldsymbol{w}} \sum_{i=1}^{N} \log p(t_i|\eta) = \sum_{i=1}^{N} (t_i - y_i)\phi(\boldsymbol{x_i})$$

GLM example: logistic regression

Cross-entropy loss (negative log-likelihood):

$$\mathcal{L} = -\log p(t|y) = -t \log y - (1-t)\log(1-y)$$

Corresponding label distribution: Bernoulli

$$\begin{aligned} p(t|y) &= y^t(1-y)^{1-t} \\ &= (1-y)\exp\{(\log \frac{y}{1-y})t\} \end{aligned}$$

We have

$$\eta = \log \frac{y}{1-y} = \psi(y)$$

Then we should choose activation:

$$a(z) = \psi^{-1}(z) = \sigma(z) \qquad \checkmark$$

GLM example: linear regression

Squared loss (negative log-likelihood):

$$\mathcal{L} = -\log p(t|y) = \frac{1}{2}(t-y)^2$$

Corresponding label distribution: Gaussian (with fixed σ, WLOG assume $\sigma = 1$)

$$p(t|y) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(t-y)^2\}$$

We have

$$\eta = y = \psi(y)$$

Then we should choose activation:

$$a(z) = \psi^{-1}(z) = z \qquad \checkmark$$

Summary

Exponential family

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\eta})} h(\boldsymbol{x}) \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(\boldsymbol{x}))$$

- Many common distributions belong to this family (Bernoulli, multinomial, Gaussian, Poisson, gamma, ...)
- Sufficient statistics for maximum-likelihood estimation
- Many useful identities stemming from $\mathcal{Z}(\boldsymbol{\eta})$
  - Moments & empirical moments, MLE as moment matching
  - Convenient way to compute the Fisher information matrix $\boldsymbol{F}_{\boldsymbol{\eta}}$
- Used to derive the generalized linear models (GLM)