DM825

Introduction to Machine Learning

Lecture 4

Model Assessment
Generalized Linear Models

Marco Chiarandini

Department of Mathematics & Computer Science
University of Southern Denmark

Outline

1. Error Estimation Methods

2. Generalized Linear Models

Outline

1. Error Estimation Methods

2. Generalized Linear Models

Loss Function in Classification

- $\mathcal{G} = \{1, \ldots, k\}$
- $p_k(\vec{x}) = \Pr(G = k \mid \vec{X} = \vec{x})$ the probability modeled
- $\hat{G}(\vec{x}) = \operatorname{argmax}_k \hat{p}_k(\vec{x})$ predicted

$$L(G, \hat{G}(\vec{x})) = I(G \neq \hat{G}(\vec{x})) \qquad \text{0–1 loss}$$

$$\begin{aligned} L(G, \hat{G}(\vec{x})) &= -2 \sum_{k=1}^{K} I(G = k) \log_2 \hat{p}_k(\vec{x}) \qquad \text{entropy} \\ &= -2 \log_2 \hat{p}_G(\vec{x}) \end{aligned}$$

Akaike Information Criterion

$$AIC = log(p(\mathcal{D} \mid \theta)) - p$$

requires an adjustment of max likelihood to account for different complexities in the models choose model with largest AIC:

computed on training set only.

Methods to Estimate Error Curves

Model selection: estimate performance in order to choose the best model

model assessment: selected a final model, estimating its prediction error on new data.

If plenty of data, divide data randomly and use:

- 50% for training
- 25% for model selection (validation)
- 25% for assessment

If less data:

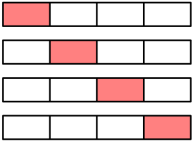
- cross validation
- Bootstrap method

Cross Validation

k-fold cross validation: k parts of m/k elements

leave k part out and use the rest of the data to train the model

(if $k = m$ then leave-one-out)



We use extra sample to estimate error $Err = E[L(Y, h(\mathbf{x}))]$ where $(Y, \vec{X})$ from joint distribution

for i from 1 to k **do**
- take out the ith part
- fit models on other $k-1$ parts
- calculate prediction error when predicting ith part

$\varphi : \{1 \ldots m\} \rightarrow \{1 \ldots k\}$ by randomization
$\hat{h}^{-i}(\vec{x})$ fitted function on data $\vec{x}$ with ith part removed

$$CV = \frac{1}{m}\sum_{i=1}^{m}(L(y^i, \hat{h}^{-\varphi(i)}(\vec{x}_i))$$

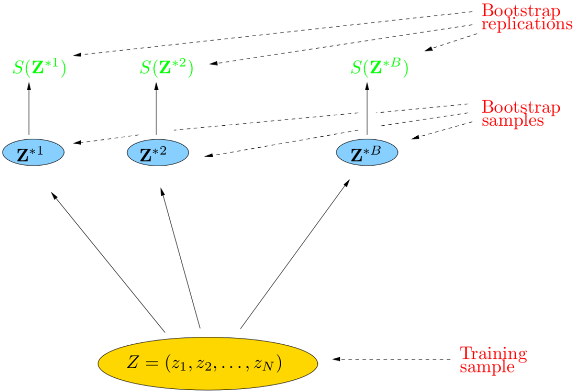
$k = 5, 10$ search $\hat{\theta}$ that minimizes CV.

Bootstrap Method

Training set $\vec{z} = (z^1, z^2, \ldots, z^m)$ and $z^i = (x^i, y^i)$

randomly draw data sets with replacement

repeat
  draw a data set
  fit the model
until $B = 100$ times ;



We can estimate any aspect of $S(\vec{z})$

$$\widehat{\mathrm{Var}}[S(\vec{z})] = \frac{1}{B-1}\sum_{b=1}^{B}(S(z^{*b}) - \bar{S}^{*})^2$$

$$\widehat{Err}_{boost} = \frac{1}{B}\frac{1}{m}\sum_{b=1}^{B}\sum_{j=1}^{m} L(y^i, \hat{h}^{*b}(x^i))$$

$\hat{h}^{*b}(x^i)$ is predicted value at $\vec{x}^i$ of model fitted on bth. There are common observations between training and test observations. To avoid this:

$$\widehat{Err}_{boost} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y^i, \hat{h}^{*b}(x^i))$$

C^{-i} is set of indices of the bootstrap samples b that do not contain observation i.

Outline

Exponential Family of Distributions

We have seen:

- regression $y \mid x; \theta \sim \mathcal{N}(\mu, \sigma^2)$
- classification $y \mid x; \theta \sim \mathrm{Bern}(\mu, \sigma^2)$

They can be shown to belong to the framework: GLM

Exponential distribution:

$$p(\vec{y} \mid \eta) = c(\vec{y})g(\vec{\eta}) \exp\{\vec{\eta}^T \vec{u}(\vec{y})\} = b(\vec{y}) \exp\{\vec{\eta}^T \vec{T}(\vec{y}) - a(\vec{\eta})\}$$

$\vec{y}$ scalar or vector, discrete or continuous

$\vec{\eta}$ canonical or natural parameters

$\vec{u}(\vec{y})$ function of $\vec{y}$

$g(\vec{\eta})$ ensures the distribution is normalized:

$$g(\vec{\eta}) \int c(\vec{y}) \exp\{\vec{\eta}^T \vec{u}(\vec{y})\} d\vec{y} = 1$$

$$c(y) = b(y)$$
$$u(y) = T(y)$$
$$g(\eta) = \frac{1}{\exp(a(\eta))}$$

Exponential Family of Distributions
Gaussian distribution

Gaussian distribution with $\sigma^2 = 1$ as an exponential distribution

$$p(y \mid \mu) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y-\mu)^2\right\}$$
$$= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}y^2\right\} \exp\left\{\mu y - \frac{1}{2}\mu^2\right\}$$

$$\eta = \mu$$
$$u(y) = y$$
$$c(y) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}y^2\right\}$$
$$g(\eta) = \exp\left\{-\frac{\mu^2}{2}\right\}$$

Exponential Family of Distributions
Gaussian distribution

Gaussian distribution as an exponential distribution

$$p(y \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}y^2\right\} \exp\left\{\frac{\mu y}{\sigma^2} - \frac{1}{2\sigma^2}\mu^2\right\}$$

$$\vec{\eta} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$$
$$\vec{u}(y) = \begin{bmatrix} y \\ y^2 \end{bmatrix}$$
$$c(y) = \frac{1}{\sqrt{2\pi}}$$
$$g(\vec{\eta}) = \sqrt{-2\eta_2} \exp\left\{\frac{\eta_1^2}{4\eta_2}\right\}$$

Exponential Family of Distributions

Bernoulli distribution

Bernoulli distribution as an exponential distribution

$$\begin{aligned} p(y \mid \mu) &= \mathrm{Bern}(y \mid \mu) = \mu^y (1-\mu)^{1-y} \\ &= \exp\{y \log \mu + (1-y)\log(1-\mu)\} \\ &= \exp\{y \log \mu + \log(1-\mu) - y\log(1-\mu)\} \\ &= (1-\mu)\exp\left\{\log\left(\frac{\mu}{1-\mu}\right) y\right\} \end{aligned}$$

exponent of log

$\eta = \log \frac{\mu}{1-\mu}$
link function

$\mu = \sigma(\eta) = \frac{1}{1+\exp(-\eta)}$
response function

$1-\mu = 1-\sigma(\eta)$

$1-\sigma(\eta) = \sigma(-\eta)$

$$p(y \mid \eta) = \sigma(-\eta)\exp(\eta y)$$

$$\begin{aligned} u(y) &= y \\ c(y) &= 1 \\ g(\eta) &= \sigma(-\eta) \end{aligned}$$

Exponential Family of Distributions

Multinomial distribution

$y \in \{1, 2, \dots k\}$ modeled as multinomial variable: $\vec{y} \mid \theta \sim \text{Multinomial}(\vec{\mu})$
$\sum_{j=1}^{k} \mu_j = 1 \rightsquigarrow \mu_1, \dots \mu_{k-1}$ independent parameters $\rightsquigarrow p(y = j \mid \vec{\mu}) = \mu_j$
and $p(y = k \mid \vec{\mu}) = \mu_k = 1 - \sum_{j=1}^{k-1} \mu_j$

$$p(\vec{y} \mid \vec{\mu}) = \Pi_{j=1}^{k} \mu_j^{x_j}$$

$$= \exp\left\{ \sum_{j=1}^{k} y_j \ln \mu_j \right\}$$

$$\vec{y} = (y_1, \dots, y_k)$$

$$p(\vec{y} \mid \vec{\eta}) = \exp(\vec{\eta}^T \vec{y})$$

$$\eta_j = \ln \mu_j, \quad \vec{\eta} = (\eta_1, \dots, \eta_m)$$

$$\vec{u}(\vec{y}) = \vec{y}$$

$$c(\vec{y}) = 1$$

$$g(\vec{\eta}) = 1$$

removing the constraint that $\sum_{j=1}^{k} \mu_j = 1$

$$\exp\left\{\sum_{j=1}^{k} y_j \ln \mu_j\right\} = \exp\left\{\sum_{j=1}^{k-1} y_j \ln \mu_j + (1 - \sum_{j=1}^{k-1} y_j) \ln(1 - \sum_{j=1}^{k-1} \mu_j)\right\}$$

$$= \exp\left\{\sum_{j=1}^{k-1} y_j \ln \frac{\mu_j}{(1 - \sum_{j=1}^{m-1} y_j)} + \ln(1 - \sum_{j=1}^{k-1} \mu_j)\right\}$$

$$\ln \frac{\mu_j}{(1 - \sum_{j=1}^{k-1} y_j)} = \eta_j$$

$$\mu_j = \frac{\exp(\eta_j)}{1 + \sum_{j}^{k-1} \exp(\eta_j)}$$ softmax function

$$p(\vec{y} \mid \vec{\eta}) = \frac{\exp(\vec{\eta}^T \vec{x})}{1 + \sum_{j=1}^{k-1} \exp(\eta_j)}$$

$$\vec{u}(\vec{y}) = \vec{y}$$

$$c(\vec{y}) = 1$$

$$g(\vec{y}) = \frac{1}{1 + \sum_{j=1}^{k-1} \exp(\eta_j)}$$

Exponential Family of Distributions

Other distributions:

- Poisson (for counting problems)
- gamma and exponential (for continuous nonnegative random variables, such as time intervals)
- beta and Dirichelet (for distributions over probabilities)

Maximum Likelihood

estimate parameter $\vec{\eta}$ in general exponential family distribution
$\mathbf{X} = (\vec{x}^1, \ldots, \vec{x}^m)$ training data

$$p(\mathbf{X} \mid \vec{\eta}) = \left(\prod_{i=1}^{m} h(\vec{x}^i) \right) g(\vec{\eta})^m \exp\left\{ \vec{\eta}^T \sum_{i=1}^{m} \vec{u}(\vec{x}^i) \right\}$$

$$-\nabla \log g(\eta_{ML}) = \frac{1}{m} \sum_{i=1}^{m} \vec{u}(\vec{x}^i)$$

Conjugate Priors

we seek a prior that is conjugate to the likelihood function such that the posterior has the same functional form as the prior

$$p(\vec{\eta} \mid \mathbf{X}, \vec{\chi}, \nu) = f(\vec{\chi}, \nu) g(\vec{\eta})^{\nu} \exp\{\nu \vec{\eta}^T \vec{\chi}\}$$

Constructing GLM

Consider a classification or a regression problem $(y, \vec{x})$. Predict y as a function of $\vec{x}$. (eg, predict number of page views in our web site based on certain features such as time of the day, advertising, etc.)

Assumptions:

1. $y \mid \vec{x}; \theta \sim \mathrm{ExpFam}(\vec{\eta})$

2. given $\vec{x}$, predict expected value of $u(y)$:
   if $u(y) = y \implies h(y) = E[y \mid \vec{x}]$

3. $\vec{\eta}$ and input $\vec{x}$ are related linearly (linear predictor):

$$\eta = \vec{\theta}^T \vec{x} \qquad (\eta_i = \vec{\theta}_i^T \vec{x})$$

Ordinary Least Squares

$$y \mid \vec{x}; \theta \sim \mathcal{N}(\mu, \sigma^2)$$

$h_{\vec{\theta}}(\vec{x}) = E[y \mid \vec{x}; \theta]$	assumption 2.
$= \mu$	because normal
$= \eta$	ass. 1 + what shown before
$= \theta^T \vec{x}$	ass. 2.

Logistic Regression

$y \mid \vec{x}; \theta \sim \mathrm{Bern}(\mu)$

$$\begin{aligned} h_{\vec{\theta}}(\vec{x}) &= E[y \mid \vec{x}; \theta] && \text{assumption 2.} \\ &= \mu && \text{because Bernoulli} \\ &= \frac{1}{1 + \exp(-\vec{\eta})} && \text{ass. 1 + what shown before} \\ &= \frac{1}{1 + \exp(-\vec{\theta}^T \vec{x})} && \text{ass. 2.} \end{aligned}$$

This answers also the question why the logistic sigmoid function was chosen

$g(\eta) = E[\vec{u}(\vec{x}); \eta]$ canonical response function

g^{-1} canonical link function

Multinomial Regression

$y \in \{1, 2, \ldots k\}$ modeled as multinomial variable:
$y \mid \vec{x}; \theta \sim \mathrm{Multinomial}(\vec{\mu})$
$\sum_{i=1}^{k} \mu_i = 1 \rightsquigarrow \mu_1, \ldots \mu_{k-1}$ independent parameters $\rightsquigarrow p(y = j \mid \vec{\mu}) = \mu_j$
and $p(y = k \mid \vec{\mu}_j) = \mu_k = 1 - \sum_{i=1}^{k-1} \mu_i$

$$p(\vec{y} \mid \vec{\mu}) = \Pi_{j=1}^{k} \mu_j^{y_i} = \frac{\exp(\vec{\eta}^T \vec{y})}{1 + \sum_{j=1}^{k-1} \exp(\eta_j)}$$

$\vec{y} = (y_1, \ldots, y_k)$

$$h_{\vec{\theta}}(\vec{x}) = E[u(\vec{y}) \mid \vec{x}; \theta] = E[y \mid \vec{x}; \theta] = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} = \begin{bmatrix} \frac{\exp(\eta_1)}{1+\sum_{i=1}^{k-1} \exp(\eta_j)} \\ \vdots \\ \frac{\exp(\eta_k)}{1+\sum_{j=1}^{k-1} \exp(\eta_j)} \end{bmatrix}$$

assumption 2.

because multinomial
ass. 1 + what shown before
estimate η by $\vec{\theta}\vec{x}$

Estimation of parameters θ via loglikelihood ℓ:

$$
\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{m} \log p(y^{(i)} | x^{(i)}; \theta) \\
&= \sum_{i=1}^{m} \log \prod_{l=1}^{k} \left(\frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \right)^{1\{y^{(i)}=l\}}
\end{aligned}
$$

and maximize by gradient ascent