Linear and logistic regression

Guillaume Obozinski

Ecole des Pontes - ParisTech

Master MVA 2014-2015

Outline

Linear regression

Generative models vs conditional models

- X is the input variable
- Y is the output variable

A **generative model** is a model of the joint distribution $p(x, y)$.

A **conditional model** is a model of the conditional distribution $p(y|x)$.

Conditional models *vs* Generative models

- CM make fewer assumptions about the data distribution
- CM require fewer parameters
- CM are typically computationally harder to learn
- CM can typically not handle missing data or latent variables

Probabilistic version of linear regression

Modeling the conditional distribution of Y given X by

$$Y \mid X \sim \mathcal{N}(\boldsymbol{w}^\top X + b, \sigma^2)$$

or equivalently $Y = \boldsymbol{w}^\top X + b + \epsilon \quad \text{with} \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$.

The offset can be ignored up to a reparameterization.

$$Y = \tilde{\boldsymbol{w}}^\top \begin{pmatrix} x \\ 1 \end{pmatrix} + \epsilon.$$

Likelihood for one pair

$$p(y_i \mid \mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{1}{2}\frac{(y_i - \boldsymbol{w}^\top \mathbf{x}_i)^2}{\sigma^2}\right)$$

Negative log-likelihood

$$-\ell(\boldsymbol{w}, \sigma^2) = -\sum_{i=1}^{n} \log p(y_i|\mathbf{x}_i) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \boldsymbol{w}^\top \mathbf{x}_i)^2}{\sigma^2}.$$

Probabilistic version of linear regression

$$\min_{\sigma^2,\boldsymbol{w}} \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \boldsymbol{w}^\top \mathbf{x}_i)^2}{\sigma^2}$$

The minimization problem in $\boldsymbol{w}$

$$\min_{\boldsymbol{w}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2$$

that we recognize as the usual linear regression, with

- $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ and
- $\boldsymbol{X}$ the design matrix with rows equal to $\mathbf{x}_i^\top$.

Optimizing over σ^2, we find:

$$\widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\boldsymbol{w}}^\top_{MLE}\mathbf{x}_i)^2$$

Logistic regression

Logistic regression

Classification setting:

$$\mathcal{X} = \mathbb{R}^p, \mathcal{Y} \in \{0, 1\}.$$

Key assumption:

$$\log \frac{\mathbb{P}(Y = 1 \mid X = \mathbf{x})}{\mathbb{P}(Y = 0 \mid X = \mathbf{x})} = \boldsymbol{w}^\top \mathbf{x}$$

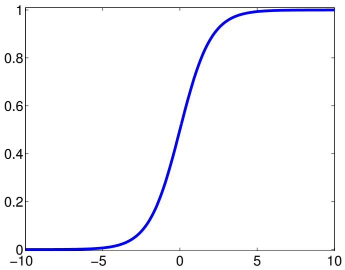
Implies that

$$\mathbb{P}(Y = 1 \mid X = \mathbf{x}) = \sigma(\boldsymbol{w}^\top \mathbf{x})$$

for

$$\sigma : z \mapsto \frac{1}{1 + e^{-z}},$$

the logistic function.



- The logistic function is part of the family of *sigmoid functions*.
- Often called "the" sigmoid function.

Properties:

$$\begin{aligned}
\forall z \in \mathbb{R}, \quad \sigma(-z) &= 1 - \sigma(z), \\
\forall z \in \mathbb{R}, \quad \sigma'(z) &= \sigma(z)(1 - \sigma(z)) \\
&= \sigma(z)\sigma(-z).
\end{aligned}$$

Likelihood for logistic regression

Let $\eta := \sigma(\boldsymbol{w}^\top \mathbf{x} + b)$. W.l.o.g. we assume $b = 0$.
By assumption: $Y|X = \mathbf{x} \ \sim \text{Ber}(\eta)$.

Likelihood

$$p(Y = y|X = \mathbf{x}) = \eta^y (1-\eta)^{1-y} = \sigma(\boldsymbol{w}^\top \mathbf{x})^y \sigma(-\boldsymbol{w}^\top \mathbf{x})^{1-y}.$$

Log-likelihood

$$\begin{aligned}
\ell(\boldsymbol{w}) &= y \log \sigma(\boldsymbol{w}^\top \mathbf{x}) + (1-y) \log \sigma(-\boldsymbol{w}^\top \mathbf{x}) \\
&= y \log \eta + (1-y) \log(1-\eta) \\
&= y \log \frac{\eta}{1-\eta} + \log(1-\eta) \\
&= y\, \boldsymbol{w}^\top \mathbf{x} + \log \sigma(-\boldsymbol{w}^\top \mathbf{x})
\end{aligned}$$

Maximizing the log-likelihood

Log-likelihood of a sample

Given an i.i.d. training set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$

$$\ell(\boldsymbol{w}) = \sum_{i=1}^{n} y_i \boldsymbol{w}^\top \mathbf{x}_i + \log \sigma(-\boldsymbol{w}^\top \mathbf{x}_i).$$

The log-likelihood is differentiable and concave.

$\Rightarrow$ Its global maxima are its stationary points.

Gradient of ℓ

$$\begin{aligned} \nabla \ell(\boldsymbol{w}) &= \sum_{i=1}^{n} y_i \mathbf{x}_i - \mathbf{x}_i \frac{\sigma(-\boldsymbol{w}^\top \mathbf{x}_i)\sigma(\boldsymbol{w}^\top \mathbf{x}_i)}{\sigma(-\boldsymbol{w}^\top \mathbf{x}_i)} \\ &= \sum_{i=1}^{n} (y_i - \eta_i)\mathbf{x}_i \qquad \text{with} \qquad \eta_i = \sigma(\boldsymbol{w}^\top \mathbf{x}_i). \end{aligned}$$

Thus, $\quad \nabla \ell(\boldsymbol{w}) = 0 \Leftrightarrow \sum_{i=1}^{n} \mathbf{x}_i (y_i - \sigma(\theta^\top \mathbf{x}_i)) = 0.$

No closed form solution !

Second order Taylor expansion

Need an iterative method to solve $\sum_{i=1}^{n} \mathbf{x}_i(y_i - \sigma(\theta^\top \mathbf{x}_i)) = 0.$

- $\rightarrow$ Gradient descent (aka steepest descent)
- $\rightarrow$ Newton's method

Hessian of ℓ

$$\begin{aligned} H\ell(\boldsymbol{w}) &= \sum_{i=1}^{n} \mathbf{x}_i(0 - \sigma'(\boldsymbol{w}^\top \mathbf{x}_i)\sigma'(-\boldsymbol{w}^\top \mathbf{x}_i)\mathbf{x}_i^\top) \\ &= \sum_{i=1}^{n} -\eta_i(1-\eta_i)\mathbf{x}_i\mathbf{x}_i^\top = -\boldsymbol{X}^\top \mathrm{Diag}(\eta_i(1-\eta_i))\boldsymbol{X} \end{aligned}$$

where $\boldsymbol{X}$ is the design matrix.
$\rightarrow$ Note that $-H\ell$ is p.s.d. $\Rightarrow \ell$ is concave.

Newton's method

Use the Taylor expansion

$$\ell(\boldsymbol{w}^t) + (\boldsymbol{w} - \boldsymbol{w}^t)^\top \nabla \ell(\boldsymbol{w}^t) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}^t)^\top H\ell(\boldsymbol{w}^t)(\boldsymbol{w} - \boldsymbol{w}^t).$$

and minimize w.r.t. $\boldsymbol{w}$. Setting $\boldsymbol{h} = \boldsymbol{w} - \boldsymbol{w}^t$, we get

$$\max_{\boldsymbol{h}} \boldsymbol{h}^\top \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}) + \frac{1}{2}\boldsymbol{h}^\top H\ell(\boldsymbol{w})\boldsymbol{h}.$$

I.e., for logistic regression, writing $\boldsymbol{D}_{\boldsymbol{\eta}} = \mathrm{Diag}\big((\eta_i(1-\eta_i))_i\big)$

$$\min_{\boldsymbol{h}} \quad \boldsymbol{h}^\top \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{\eta}) - \frac{1}{2}\boldsymbol{h}^\top \boldsymbol{X}^\top \boldsymbol{D}_{\boldsymbol{\eta}} \boldsymbol{X} \boldsymbol{h}$$

Modified normal equations

$$\boldsymbol{X}^\top \boldsymbol{D}_{\boldsymbol{\eta}} \boldsymbol{X}\, \boldsymbol{h} - \boldsymbol{X}^\top \tilde{\boldsymbol{y}} \qquad \text{with} \qquad \tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{\eta}.$$

Iterative Reweighted Least Squares (IRLS)

Assuming $\boldsymbol{X}^\top \boldsymbol{D}_{\boldsymbol{\eta}} \boldsymbol{X}$ is invertible, the algorithm takes the form

$$\boldsymbol{w}^{(t+1)} \leftarrow \boldsymbol{w}^{(t)} + (\boldsymbol{X}^\top \boldsymbol{D}_{\boldsymbol{\eta}^{(t)}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{\eta}^{(t)}).$$

This is called iterative reweighted least squares because each step is equivalent to solving the reweighted least squares problem:

$$\frac{1}{2} \sum_{i=1}^{n} \frac{1}{\tau_i^2} (\mathbf{x}_i^\top \boldsymbol{h} - \check{y}_i)^2$$

with

$$\tau_i^2 = \frac{1}{\eta_i^{(t)} (1 - \eta_i^{(t)})} \qquad \text{and} \qquad \check{y}_i = \tau_i^2 (y_i - \eta_i^{(t)}).$$

Alternate formulation of logistic regression

If $y \in \{-1, 1\}$, then

$$\mathbb{P}(Y = y | X = \mathbf{x}) = \sigma(y\, \boldsymbol{w}^\top \mathbf{x})$$

Log-likelihood

$$\ell(\boldsymbol{w}) = \log \sigma(y\boldsymbol{w}^\top \mathbf{x}) = -\log\big(1 + \exp(-y\boldsymbol{w}^\top x)\big)$$

Log-likelihood for a training set

$$\ell(\boldsymbol{w}) = -\sum_{i=1}^{n} \log\big(1 + \exp(-y_i \boldsymbol{w}^\top x_i)\big)$$

Fisher discriminant analysis

Generative classification

$X \in \mathbb{R}^p$ and $Y \in \{0, 1\}$. Instead of modeling directly $p(y \mid \mathbf{x})$ model $p(y)$ and $p(\mathbf{x} \mid y)$ and deduce $p(y \mid \mathbf{x})$ using Bayes rule.
In classification $\mathbb{P}(Y = 1 \mid X = \mathbf{x}) =$

$$\frac{\mathbb{P}(X = \mathbf{x} \mid Y = 1)\,\mathbb{P}(Y = 1)}{\mathbb{P}(X = \mathbf{x} \mid Y = 1)\,\mathbb{P}(Y = 1) + \mathbb{P}(X = \mathbf{x} \mid Y = 0)\,\mathbb{P}(Y = 0)}$$

For example one can assume

- $\mathbb{P}(Y = 1) = \pi$
- $\mathbb{P}(X = \mathbf{x} \mid Y = 1) \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
- $\mathbb{P}(X = \mathbf{x} \mid Y = 0) \sim \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.

Fisher's discriminant aka Linear Discriminant Analysis (LDA)

Previous model with the constraint $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_0 = \mathbf{\Sigma}$. Given a training set, the different model parameters can be estimated using the maximum likelihood principle, which leads to

$$(\widehat{\pi}, \widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_0, \widehat{\mathbf{\Sigma}}_1, \widehat{\mathbf{\Sigma}}_0).$$