Regression Methods

Linear Regression and Logistic Regression:

definitions, and a common property

CMU, 2004 fall, Andrew Moore, HW2, pr. 4

Linear Regression and Logistic Regression: Definitions

Given an input vector X, linear regression models a real-valued output Y as

$$Y|X \sim Normal(\mu(X), \sigma^2),$$

where $\mu(X) = \beta^\top X = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

Given an input vector X, logistic regression models a binary output Y by

$$Y|X \sim Bernoulli(\theta(X)),$$

where the Bernoulli parameter is related to $\beta^\top X$ by the *logit* transformation

$$logit\,(\theta(X)) \stackrel{def.}{=} \log \frac{\theta(X)}{1-\theta(X)} = \beta^\top X.$$

a. For each of the two regression models defined above, write the log likelihood function and its gradient with respect to the parameter vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$.

Answer:

For *linear regression*, we can write the log likelihood function as:

$$
\begin{aligned}
LL(\beta) &= \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \mu(x_i))^2}{2\sigma^2}\right)\right) \\
&= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \beta^\top x_i)^2}{2\sigma^2}\right)\right) \\
&= -n\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta^\top x_i)^2 \\
&= -n\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta^\top x_i)^\top (y_i - \beta^\top x_i).
\end{aligned}
$$

Therefore, its gradient is:

$$
\nabla_\beta LL(\beta) = \sum_{i=1}^{n}(y_i - \beta^\top x_i)x_i
$$

For *logistic regression*:

$$\log \frac{\theta(X)}{1-\theta(X)} = \beta^\top X \Leftrightarrow e^{\beta^\top X} = \frac{\theta(X)}{1-\theta(X)} \Leftrightarrow e^{\beta^\top X} = \theta(X)(1+e^{\beta^\top X})$$

Therefore,

$$\theta(X) = \frac{e^{\beta^\top X}}{1+e^{\beta^\top X}} = \frac{1}{1+e^{-\beta^\top X}} \textbf{ and } 1-\theta(X) = \frac{1}{1+e^{\beta^\top X}}.$$

Note that $Y|X \sim Bernoulli(\theta(X))$ **means that**

$$P(Y=1|X) = \theta(X) \textbf{ and } P(Y=0|X) = 1-\theta(X),$$

which can be equivalently written as

$$P(Y=y|X) = \theta(X)^y(1-\theta(X))^{1-y} \text{ for all } y \in \{0,1\}.$$

So, in this case the log likelihood function is:

$$\begin{aligned} LL(\beta) &= \log\left(\prod_{i=1}^{n}\{\theta(x_i)^{y_i}(1-\theta(x_i))^{1-y_i}\}\right) \\ &= \sum_{i=1}^{n}\{y_i \log\theta(x_i) + (1-y_i)\log(1-\theta(x_i))\} \\ &= \sum_{i=1}^{n}\{y_i(\beta^\top x_i + \log(1-\theta(x_i)) + (1-y_i)\log(1-\theta(x_i))\} \\ &= \sum_{i=1}^{n}\{y_i(\beta^\top x_i) - \log(1+e^{\beta^\top x_i})\} \end{aligned}$$

And therefore,

$$\nabla_\beta LL(\beta) = \sum_{i=1}^{n}\left(y_i x_i - \frac{e^{\beta^\top x_i}}{1+e^{\beta^\top x_i}}x_i\right) = \sum_{i=1}^{n}(y_i - \theta(x_i))x_i$$

Remark

Actually, in the above solutions the full log likelihood function should look like the following first:

$$
\begin{aligned}
\text{log-likelihood} &= \log \prod_{i=1}^{n} p(x_i, y_i) \\
&= \log \prod_{i=1}^{n} (p_{Y|X}(y_i|x_i)\, p_X(x_i)) \\
&= \log \left(\left(\prod_{i=1}^{n} p_{Y|X}(y_i|x_i) \right) \cdot \left(\prod_{i=1}^{n} p_X(x_i) \right) \right) \\
&= \log \prod_{i=1}^{n} p_{Y|X}(y_i|x_i) + \log \prod_{i=1}^{n} p_X(x_i) \\
&= LL + LL_x
\end{aligned}
$$

Because LL_x does not depend on the parameter β, when doing MLE we could just consider maximizing LL.

b. Show that for each of the two regression models above, at the MLE $\hat{\beta}$ has the following property:

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} E[Y|X = x_i, \beta = \hat{\beta}] x_i.$$

Answer:

For *linear regression*:

$$\nabla_\beta LL(\beta) = 0 \Rightarrow \sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} (\hat{\beta}^\top x_i) x_i.$$

Since $Y|X \sim Normal(\mu(X), \sigma^2)$**,**

$$E[Y|X = x_i, \beta = \hat{\beta}] = \mu(x_i) = \hat{\beta}^\top x_i.$$

So $\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} E[Y|X = x_i, \beta = \hat{\beta}] x_i$**.**

For *logistic regression*:

$$\nabla_\beta LL(\beta) = 0 \Rightarrow \sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} \theta(x_i) x_i.$$

Since $Y|X \sim Bernoulli(\theta(X))$**,**

$$E[Y|X = x_i, \beta = \hat{\beta}] = \theta(x_i) = \frac{e^{\hat{\beta}^\top x_i}}{1 + e^{\hat{\beta}^\top x_i}}.$$

So $\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} E[Y|X = x_i, \beta = \hat{\beta}] x_i$**.**

Linear Regression with only one parameter; MLE and MAP estimation

CMU, 2012 fall, Tom Mitchell, Ziv Bar-Joseph, midterm, pr. 3

Consider real-valued variables X and Y. The Y variable is generated, conditional on X, from the following process:

$$\varepsilon \sim N(0, \sigma^2)$$
$$Y = aX + \varepsilon,$$

where every ε is an independent variable, called a *noise* term, which is drawn from a Gaussian distribution with mean 0, and standard deviation σ.

This is a one-feature *linear regression* model, where a is the only weight parameter.

The conditional probability of Y has the distribution $p(Y|X,a) \sim N(aX, \sigma^2)$, so it can be written as

$$p(Y|X,a) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y - aX)^2\right)$$

MLE estimation

a. Assume we have a training dataset of n pairs (X_i, Y_i) for $i = 1, \ldots, n$, and σ is known. Which ones of the following equations correctly represent the maximum likelihood problem for estimating a? Say *yes* or *no* to each one. More than one of them should have the answer *yes*.

i. $\arg\max_a \sum_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$

ii. $\arg\max_a \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$

iii. $\arg\max_a \sum_i \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$

iv. $\arg\max_a \prod_i \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$

v. $\arg\max_a \sum_i (Y_i - aX_i)^2$

vi. $\mathrm{argmin}_a \sum_i (Y_i - aX_i)^2$

Answer:

$$
\begin{aligned}
L_D(a) &\stackrel{def.}{=} p(Y_1, \ldots, Y_n | a) = p(Y_1, \ldots, Y_n | X_1, \ldots, X_n, a) \\
&\stackrel{i.i.d.}{=} \prod_{i=1}^{n} p(Y_i | X_i, a) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)
\end{aligned}
$$

Therefore

$$
\begin{aligned}
a_{MLE} &\stackrel{def.}{=} \arg\max_a L_D(a) = \arg\max_a \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right) && (ii.) \\
&= \arg\max_a \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right) = \arg\max_a \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - aX_i)^2\right) \\
&= \arg\max_a \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right) && (iv.) \\
&= \arg\max_a \ln \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right) = \arg\max_a \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(Y_i - aX_i)^2 \\
&= \arg\max_a -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - aX_i)^2 = \arg\min_a \sum_{i=1}^{n} (Y_i - aX_i)^2 && (vi.)
\end{aligned}
$$

b. Derive the maximum likelihood estimate of the parameter a in terms of the training example X_i's and Y_i's. We recommend you start with the simplest form of the problem you found above.

Answer:

$$\begin{aligned} a_{MLE} &= \arg\min_a \sum_{i=1}^{n} (Y_i - aX_i)^2 = \arg\min_a \left(a^2 \sum_{i=1}^{n} X_i^2 - 2a \sum_{i=1}^{n} X_i Y_i + \sum_{i=1}^{n} Y_i^2 \right) \\ &= -\frac{-2 \sum_{i=1}^{n} X_i Y_i}{2 \sum_{i=1}^{n} X_i^2} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \end{aligned}$$

MAP estimation

Let's put a prior on a. Assume $a \sim N(0, \lambda^2)$, so

$$p(a|\lambda) = \frac{1}{\sqrt{2\pi}\lambda} \exp\left(-\frac{1}{2\lambda^2} a^2\right)$$

The posterior probability of a is

$$p(a|Y_1, \ldots, Y_n, X_1, \ldots, X_n, \lambda) = \frac{p(Y_1, \ldots, Y_n|X_1, \ldots, X_n, a)\, p(a|\lambda)}{\int_{a'} p(Y_1, \ldots, Y_n|X_1, \ldots, X_n, a')\, p(a'|\lambda)\, da'}$$

We can ignore the denominator when doing MAP estimation.

c. Assume $\sigma = 1$, and a fixed prior parameter λ. Solve for the MAP estimate of a,

$$\underset{a}{\operatorname{argmax}} [\ln p(Y_1, \ldots, Y_n|X_1, \ldots, X_n, a) + \ln p(a|\lambda)]$$

Your solution should be in terms of X_i's, Y_i's, and λ.

Answer:

$$
\begin{aligned}
& p(Y_1, \ldots, Y_n | X_1, \ldots, X_n, a) \cdot p(a|\lambda) \\
= \quad & \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} (Y_i - aX_i)^2 \right) \right) \cdot \frac{1}{\sqrt{2\pi}\lambda} \exp\left(-\frac{a^2}{2\lambda^2} \right) \\
\stackrel{\sigma=1}{=} \quad & \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} (Y_i - aX_i)^2 \right) \right) \cdot \frac{1}{\sqrt{2\pi}\lambda} \exp\left(-\frac{a^2}{2\lambda^2} \right)
\end{aligned}
$$

Therefore the MAP optimization problem is

$$
\begin{aligned}
& \arg\max_a \left(n \, \ln \frac{1}{\sqrt{2\pi}} - \frac{1}{2} \sum_{i=1}^{n} (Y_i - aX_i)^2 + \ln \frac{1}{\sqrt{2\pi}\lambda} - \frac{1}{2\lambda^2} a^2 \right) \\
= \quad & \arg\max_a \left(-\frac{1}{2} \sum_{i=1}^{n} (Y_i - aX_i)^2 - \frac{1}{2\lambda^2} a^2 \right) \\
= \quad & \arg\min_a \left(\sum_{i=1}^{n} (Y_i - aX_i)^2 + \frac{a^2}{\lambda^2} \right) = \arg\min_a \left(a^2 \left(\sum_{i=1}^{n} X_i^2 + \frac{1}{\lambda^2} \right) - 2a \sum_{i=1}^{n} X_i Y_i + \sum_{i=1}^{n} Y_i^2 \right)
\end{aligned}
$$

$$
\Rightarrow a_{MAP} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2 + \dfrac{1}{\lambda^2}}
$$

d. Under the following conditions, how do the prior and conditional likelihood curves change? Do a^{MLE} and a^{MAP} become closer together, or further apart?

	$p(a\|\lambda)$ **prior probability: wider, narrower, or same?**	$p(Y_1, \ldots, Y_n\|X_1, \ldots, X_n, a)$ **conditional likelihood: wider, narrower, or same?**	$\|a^{MLE} - a^{MAP}\|$ **increase or decrease?**
As $\lambda \to \infty$			
As $\lambda \to 0$			
More data: as $n \to \infty$ **(fixed** λ**)**			

Answer:

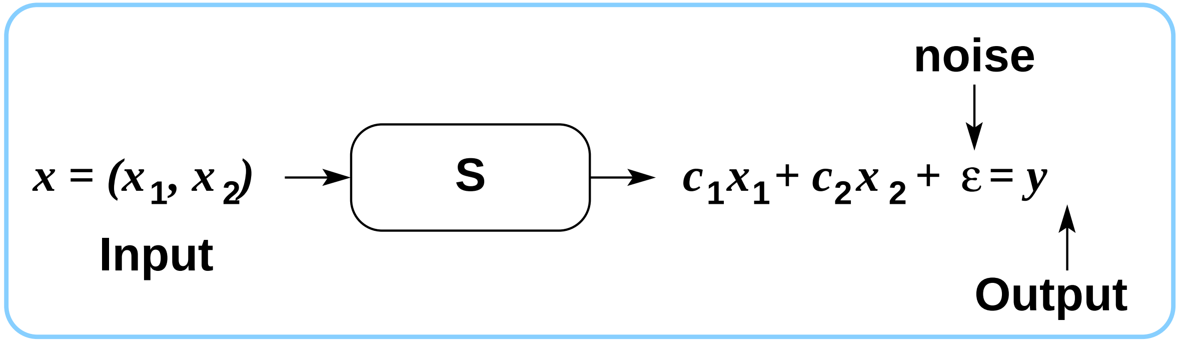
	$p(a\|\lambda)$ **prior probability: wider, narrower, or same?**	$p(Y_1, \ldots, Y_n\|X_1, \ldots, X_n, a)$ **conditional likelihood: wider, narrower, or same?**	$\|a^{MLE} - a^{MAP}\|$ **increase or decrease?**
As $\lambda \to \infty$	**wider**	**same**	**decrease**
As $\lambda \to 0$	**narrower**	**same**	**increase**
More data: as $n \to \infty$ **(fixed** λ**)**	**same**	**narrower**	**decrease**

Linear Regression in $\mathbb{R}^2$

[**without “intercept” term**]

with either Gaussian or Laplace noise

CMU, 2009 fall, Carlos Guestrin, HW3, pr. 1.5.2
CMU, 2012 fall, Eric Xing, Aarti Singh, HW1, pr. 2



This figure shows a system S which takes two inputs x_1, x_2 and outputs a linear combination of those two inputs, $c_1x_1 + c_2x_2$, where c_1 and c_2 are two unknown real numbers.

The device you use to measure the output of S, i.e., $c_1x_1 + c_2x_2$, introduces an additive error ε, which is a random variable following some distribution. Thus, the output y that you observe is given by equation (1):

$$y = c_1x_1 + c_2x_2 + \varepsilon \tag{1}$$

Assume that you have $n > 2$ instances $\langle x_{j1}, x_{j2}, y_j \rangle_{j=1,\ldots,n}$ or equivalently $\langle x_j, y_j \rangle_{j=1,\ldots,n}$, where $x_j \stackrel{not.}{=} [x_{j1}, x_{j2}]$. In other words, having n measurements in your hands is equivalent to having n equations of the following form: $y_j = c_1x_{j1} + c_2x_{j2} + \varepsilon_j$, $j = 1, \ldots, n$.

The *goal* is to estimate c_1 and c_2 from those measurements using the maximum likelihood.

a. Assume that the ε_i for $i = 1, \ldots, n$ are i.i.d. Gaussian random variables with zero mean and variance σ^2.

Compute the loglikelihood function and use it to prove that the maximum likelihood estimate $c^* = [c_1^*, c_2^*]$ is the solution of a least squares approximation problem. Find the solution of the least squares problem.

Answer:

$\varepsilon_i = y_i - (c_1 x_{i1} + c_2 x_{i2}) \sim \mathcal{N}(0, \sigma^2)$. **Therefore** $y_i \sim \mathcal{N}(c_1 x_{i1} + c_2 x_{i2}, \sigma^2)$. **Since the noise are i.i.d., the likelihood function is given by**

$$L(c_1, c_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - c_1 x_{i1} - c_2 x_{i2})^2}{2\sigma^2} \right).$$

Taking the logarithm, we get the loglikelihood function:

$$l(c_1, c_2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - c_1 x_{i1} - c_2 x_{i2})^2.$$

Let $y \in \mathbb{R}^n$ **be the vector containing the measurements,** X **the** $n \times 2$ **matrix with** $X_{ij} = x_{ij}$ **and** $c = [c_1, c_2]^\top$**, then we are trying to minimize** $||y - Xc||_2^2$ **resulting in a *solution*** $c = (X^\top X)^{-1} X^\top y$**.**

b. Assume that the ε_i for $i = 1, \ldots, n$ are independent Gaussian random variables with zero mean and variance $Var(\varepsilon_i) = \sigma_i^2$.

Compute the loglikelihood function and find $c^* = [c_1^*, c_2^*]$ which maximizes it, i.e., the MLE.

Answer:

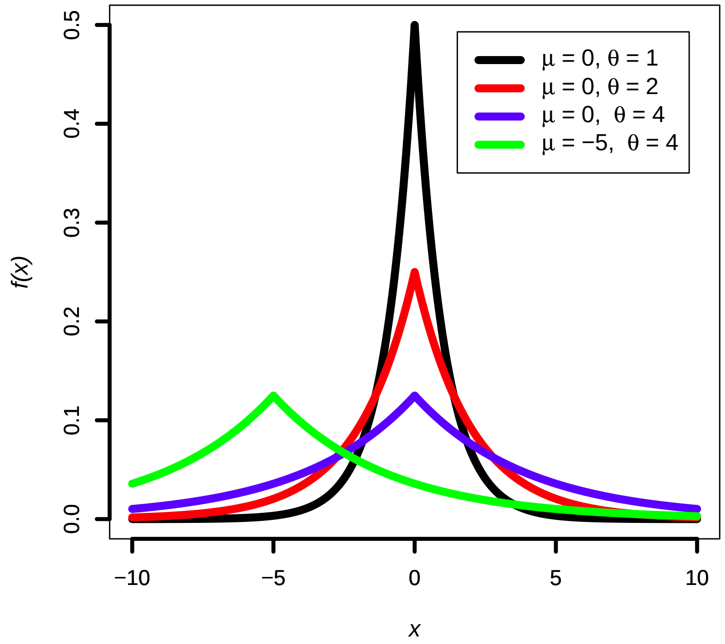
$\varepsilon_i = y_i - (c_1 x_{i1} + c_2 x_{i2}) \sim \mathcal{N}(0, \sigma_i^2)$.
Similar as before,

$$l(c_1, c_2) = -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(y_i - c_1 x_{i1} - c_2 x_{i2})^2}{2\sigma_i^2}.$$

Now we are trying to minimize $||W(y - Xc)||_2^2$, where W is a diagonal matrix, with $w_{ii} = \dfrac{1}{\sigma_i}$, resulting the solution $c = (X^\top W^\top W X)^{-1} X^\top W^\top W y$.

c. Assume that ε_i for $i = 1, \ldots, n$ has density $f_{\varepsilon_i}(x) = f(x) = \frac{1}{2b} exp(-\frac{|x|}{b})$. In other words, our noise is i.i.d. following a Laplace distribution with location parameter $\mu = 0$ and scale parameter b. Compute the loglikelihood function under this noise model and explain why this model leads to more robust solutions.

Laplace p.d.f.



Answer:

$$l(c_1, c_2) = -n \log(2b) - \sum_{i=1}^{n} ||y - Xc||_1^2.$$

It is prepared to see higher values of residuals because it has a larger tail [LC: than the Gaussian]. Thus it is more robust to noise and outliers.