Regression Methods

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

Answer:

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

Therefore

$$a_{MLE} \overset{def.}{=} \arg\max_a L_D(a) = \arg\max_a \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right) \qquad (ii.)$$

$$= \arg\max_a \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right) = \arg\max_a \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$$

$$= \arg\max_a \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right) \qquad (iv.)$$

$$= \arg\max_a \ln \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right) = \arg\max_a \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(Y_i - aX_i)^2$$

$$= \arg\max_a -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - aX_i)^2 = \arg\min_a \sum_{i=1}^{n} (Y_i - aX_i)^2 \qquad (vi.)$$

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,

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

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

Answer:

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

Therefore the MAP optimization problem is

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

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

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:

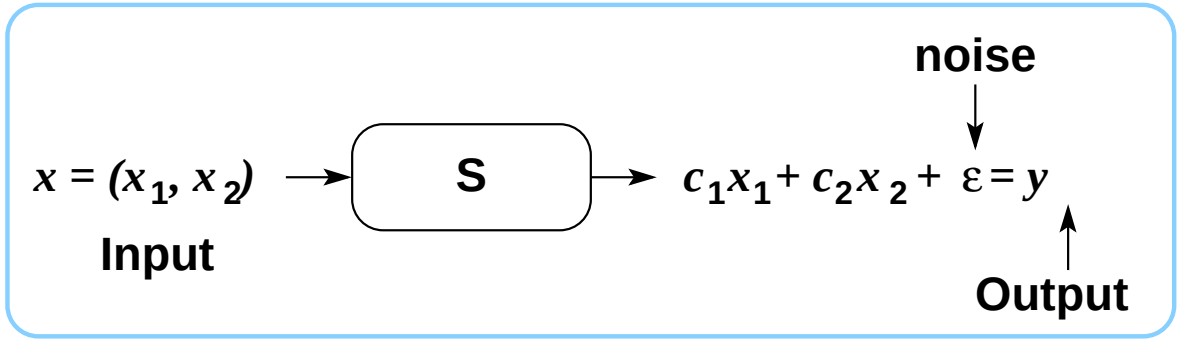
	$p(a\|\lambda)$ **prior probability: wider, narrower, or same?**	$p(Y_1, \ldots, Y_n \| X_1, \ldots, X_n, a)$ **conditional likelihood: wider, narrower, or same?**	$\|a^{MLE} - a^{MAP}\|$ **increase or decrease?**
As $\lambda \to \infty$	**wider**	**same**	**decrease**
As $\lambda \to 0$	**narrower**	**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 $x_j \overset{not.}{=} [x_{j1}, x_{j2}]$. 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,

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

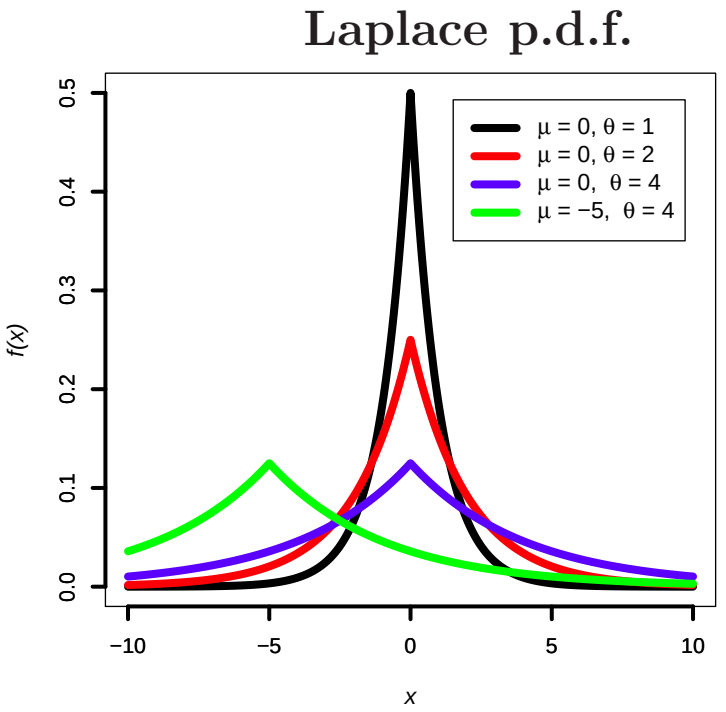
Now we are trying to minimize $||W(y - Xc)||_2^2$, where W is a diagonal matrix, with $w_{ii} = \frac{1}{\sigma_i}$, resulting the solution $c = (X^\top W^\top W X)^{-1} X^\top W^\top W y$.

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.

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.



Logistic Regression: introductory issues

Stanford, 2016 spring, Chris Piech,

Introduction to Probability for Computer Scientists course,

lecture notes #40

You are given n i.i.d. training datapoints $(x^{(1)}, y^{(1)})$, $(x^{(2)}, y^{(2)}), \ldots,$ $(x^{(n)}, y^{(n)})$, where each vector $x^{(i)}$ has d features / attributes. Here we will assume that $y^{(i)} \in \{0, 1\}$ for $i = 1, \ldots, n$.

Logistic Regression **is a classification algorithm that works by trying to learn a function that approximates $P(Y|X)$. It makes the *central assumption* that $P(Y|X)$ can be approximated as a *sigmoid function* (also called *logistic function*) applied to a linear combination of input features.**

Mathematically, for a single training datapoint (x, y) Logistic Regression assumes:

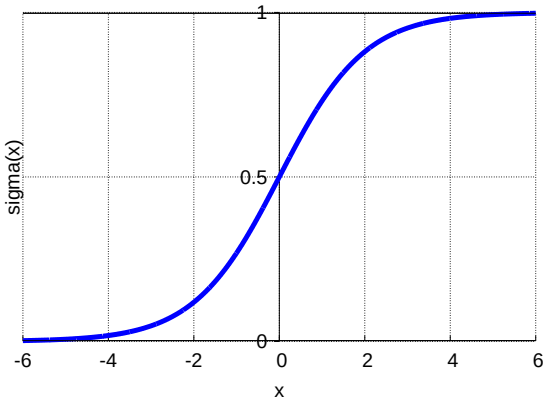
$P(Y = 1|X = x) = \sigma(z)$ **and, equivalently**

$P(Y = 0|X = x) = 1 - \sigma(z)$, **where**

$$\sigma(z) \stackrel{def.}{=} \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z}, \text{ with}$$

$$z \stackrel{not.}{=} w_0 + \sum_{i=1}^{d} w_i x_i = w \cdot x \text{ and}$$

$$w \stackrel{not.}{=} (w_0, w_1, \ldots, w_d) \in \mathbb{R}^{d+1}, \text{ assuming } x_0 = 1.$$



Starting from the above formulas for the probability of $Y|X$, we can create an algorithm that selects values of w that maximize that probability for all the training data.

a. Prove that the conditional log-likelihood of all the training data (under the Logistic Regression assumption) is:

$$L(w) = \sum_{i=1}^{n} \left(y^{(i)} \ln \sigma(w \cdot x^{(i)}) + (1 - y^{(i)}) \ln(1 - \sigma(w \cdot x^{(i)})) \right). \tag{2}$$

Solution:

To start, here is a super slick way of writing the conditional probability of one datapoint:

$$P(Y = y|X = x) = \sigma(w \cdot x)^y \left[1 - \sigma(w \cdot x)\right]^{1-y} \textbf{ assuming } y \in \{0, 1\}.$$

Since each datapoint is independent, the conditional probability of all the data is:

$$\prod_{i=1}^{n} P(Y = y^{(i)}|X = x^{(i)}) = \prod_{i=1}^{n} \sigma(w \cdot x^{(i)})^{y^{(i)}} \left[1 - \sigma(w \cdot x^{(i)})\right]^{1-y^{(i)}}. \tag{3}$$

If you apply $\ln$ **to this function, you get the reported conditional log-likelihood for Logistic Regression.**

Note

[from CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW2, pr. 4]

Actually, the *full log-likelihood* function is:

$$
\begin{aligned}
\textit{log-likelihood} &= \ln \prod_{i=1}^{n} P(x^{(i)}, y^{(i)}) \\
&= \ln \prod_{i=1}^{n} (P_{Y|X}(y^{(i)}|x^{(i)})\, P_X(x^{(i)})) \\
&= \ln \left(\left(\prod_{i=1}^{n} P_{Y|X}(y^{(i)}|x^{(i)}) \right) \cdot \left(\prod_{i=1}^{n} P_X(x^{(i)}) \right) \right) \\
&= \ln \prod_{i=1}^{n} P_{Y|X}(y^{(i)}|x^{(i)}) + \ln \prod_{i=1}^{n} P_X(x^{(i)}) \\
&= L(w) + L_x.
\end{aligned}
$$

Because L_x does not depend on the parameter w, when doing MLE we could just consider maximizing $L(w)$.

Comment

Starting from the expression (2) for the conditional log-likelihood function, we simply need to choose the values of w that maximize it.

Unlike in other situations, here there is no *closed form* way to calculate w. Instead we will choose it using *optimization*, so we will employ an algorithm called *gradient ascent*. That algorithm claims that if you continuously take small steps in the direction of your gradient, you will eventually make it to a local maxima. In the case of Logistic Regression you can prove that the result will always be *a global maxima*.

The small step that we continually take given the training dataset can be calculated as:

$$w_j^{new} = w_j^{old} + \eta \frac{\partial}{\partial w_j^{old}} L(w^{old}),$$

were η is the magnitude of the step size ("learning rate") that we take.

b. Show that the partial derivative of the conditional log-likelihood function with respect to each parameter w_j is:

$$\frac{\partial}{\partial w_j} L(w) = \sum_{i=1}^{n} [y^{(i)} - \underbrace{\sigma(w \cdot x^{(i)})}_{P(Y=1|X=x;w)}] x_j^{(i)} \text{ for } j = 0, 1, \ldots, d. \tag{4}$$

***Hint*: You may use the following property for the derivative of σ with respect to its inputs:**

$$\frac{\partial}{\partial z} \sigma(z) = \sigma(z)[1 - \sigma(z)] \text{ for } \forall z \in \mathbb{R}.$$

Solution

The partial derivative of the conditional log-likelihood for *only* one datapoint (x, y) w.r.t. the w_j component is computed as follows:

$$\begin{aligned}
&\frac{\partial}{\partial w_j} \ln[\sigma(w \cdot x)^y \, [1 - \sigma(w \cdot x)]^{1-y}] \\
&= \frac{\partial}{\partial w_j} y \ln \sigma(w \cdot x) + \frac{\partial}{\partial w_j} (1 - y) \ln[1 - \sigma(w \cdot x)] \\
&= \left[\frac{y}{\sigma(w \cdot x)} - \frac{1 - y}{1 - \sigma(w \cdot x)} \right] \frac{\partial}{\partial w_j} \sigma(w \cdot x) \\
&= \left[\frac{y}{\sigma(w \cdot x)} - \frac{1 - y}{1 - \sigma(w \cdot x)} \right] \sigma(w \cdot x)[1 - \sigma(w \cdot x)] x_j \\
&= \frac{y - \sigma(w \cdot x)}{\sigma(w \cdot x)[1 - \sigma(w \cdot x)]} \sigma(w \cdot x)[1 - \sigma(w \cdot x)] x_j \\
&= [y - \sigma(w \cdot x)] x_j . \qquad (5)
\end{aligned}$$

Because the derivative of sums is the sum of derivatives, the partial derivative of the conditional log-verosimility function w.r.t. w_j is simply the sum o this term for each training datapoint. More exactly, after applying the $\ln$ function to the (3) equality, and then calculating its partial derivative w.r.t. w_j (for $j \in \{0, 1, \ldots, d\}$) we will get the (4) result, due to the (5) relation proven above.

Another solution

[based on CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW2, pr. 4]

Starting from (2), the conditional log-likelihood function can be further written as:

$$\begin{aligned} L(w) &= \sum_{i=1}^{n} \{y_i \ln \sigma(w \cdot x_i) + (1 - y_i) \ln(1 - \sigma(w \cdot x_i))\} \\ &= \sum_{i=1}^{n} \{y_i \ln \frac{\sigma(w \cdot x_i)}{1 - \sigma(w \cdot x_i)} + \ln(1 - \sigma(w \cdot x_i))\} \overset{(*)}{=} \sum_{i=1}^{n} \{y_i (w \cdot x_i) - \ln(1 + e^{w \cdot x_i})\} \end{aligned} \quad (6)$$

And therefore, the gradient vector for $L(w)$ can be written as:

$$\nabla_w L(w) = \sum_{i=1}^{n} \left(y_i x_i - \frac{e^{w \cdot x_i}}{1 + e^{w \cdot x_i}} x_i \right) = \sum_{i=1}^{n} (y_i - \sigma(w \cdot x_i)) x_i$$

(*): We used the fact that the sigmoidal function $\sigma : \mathbb{R} \to (0, 1)$ is bijective, and its inverse function is:

$$\sigma^{-1} : (0, 1) \to \mathbb{R}, \text{ defined by } \sigma^{-1}(z) = \ln \frac{z}{1 - z}.$$

Some people call σ^{-1} the *logit* function.

Logistic Regression is not affected by variable duplication

CMU, 2011 spring, Tom Mitchell, midterm, pr. 5.3

Consider a binary classification problem with variable $X_1 \in \{0, 1\}$ and label $Y \in \{0, 1\}$.

We have a training set D_1 made of n examples: $D_1 = \{(x_1^1, y^1), \ldots, (x_1^n, y^n)\}$. Suppose we generate another training set D_2 of n examples, $D_2 = \{(x_1^1, x_2^1, y^1), \ldots, (x_1^n, x_2^n, y^n)\}$, where in each example x_1 and y are the same as in D_1 and then x_2 is a duplicate of x_1.

Now we learn a logistic regression from D_1, which should contain two parameters: w_0 and w_1; we also learn another logistic regression from D_2, which should have three parameters: w_0, w_1 and w_2.

First, write down the training rule (maximum conditional likelihood estimation) we use to estimate the parameters (w_0, w_1) and (w_0, w_1, w_2) from data. Then, given the training rule, what is the relationship between (w_0, w_1) and (w_0, w_1, w_2) we estimated from D_1 and D_2? Use this fact to argue whether or not the logistic regression will suffer from having an additional duplicate variable X_2.

Answer

The training rule for (w_0, w_1) aims to maximize the following log-likehood function:

$$\ln \prod_{i=1}^{n} P(Y^l | X_1^l, w_0, w_1) \overset{(6)}{=} \sum_{i=1}^{n} Y^l (w_0 + w_1 X_1^l) - \ln(1 + \exp(w_0 + w_1 X_1^l)).$$

Similarly, the training rule for (w_0', w_1', w_2') aims to maximize

$$\ln \prod_{i=1}^{n} P(Y^l | X_1^l, w_0', w_1', w_2') \overset{(6)}{=} \sum_{i=1}^{n} Y^l (w_0' + w_1' X_1^l + w_2' X_2^l) - \ln(1 + \exp(w_0' + w_1' X_1^l + w_2' X_2^l))$$

$$= \sum_{i=1}^{n} Y^l (w_0' + (w_1' + w_2') X_1^l) - \ln(1 + \exp(w_0' + (w_1' + w_2') X_1^l)),$$

which is basically the same as [the log-likelihood function for deriving] the training rule for (w_0, w_1), with the substitutions $w_0 = w_0'$ and $w_1 = w_1' + w_2'$.

These substitutions express the relationship between the sets of parameters (w_0, w_1) and (w_0', w_1', w_2') that we estimate from the training sets D_1 and respectively D_2.

Therefore, logistic regression will simply split the weight w_1 into $w_1' + w_2' = w_1$ when facing the duplicated variable $X_2 = X_1$.

Multi-class Logistic Regression with L_2 regularization

CMU, 2012 fall, Tom Mitchell, Ziv Bar-Joseph, HW2, pr. 2

We can easily extend the binary Logistic Regression model to handle multi-class classification. Let's assume that we have K different classes, and the posterior probability for class k is given by:

$$P(Y = k|X = x) = \frac{\exp(w_k \cdot x)}{1 + \sum_{t=1}^{K-1} \exp(w_t \cdot x)} \text{ for } k = 1, \ldots, K-1$$

$$P(Y = K|X = x) = \frac{1}{1 + \sum_{t=1}^{K-1} \exp(w_t \cdot x)},$$

where x and w_t for $t = 1, \ldots, K$ are d-dimensional vectors. Notice that we ignored the components w_{t0} in order to simplify the expression.

Our *goal* is to estimate the weights w_t using the gradient ascent optimization method. We will also define *priors* on the parameters to avoid overfitting and very large weights.

a. Assume that you are given a $n \times d$ training matrix, where n is the number of training examples, and d is the number of attributes / dimensions. Please explicitly write down the log-likelihood function, $L(w_1, \ldots, w_K)$ with L_2 regularization on the weights. Show your steps.

***Hint*: You can simplify the multi-class logistic regression expression above by introducing a fixed parameter vector $w_K = 0$ (the d-dimensional vector made entirely of $0'$s).**

b. Note that like for the binary classification case there is not a *closed form* solution to maximize the log-conditional likelihood, $L(w_1, \ldots, w_K)$, with respect to w_k. However, we can still find the solution with the *gradient ascent* method, by using partial derivatives. Derive the expression for the i-th component in the vector gradient $L(w_1, \ldots, w_K)$ with respect to w_i, which is the partial derivative of $L(w_1, \ldots, w_K)$ with respect to w_i.

c. Beginning with the initial weights of 0, write down the *update rule* for w_k, using ν for the step size. Will the solution converge to a global maximum?

Answer

a. Let $1_{\{lk\}}$ be an indicator function, where $1_{\{lk\}} = 1$ if $Y_l = k$, otherwise $1_{\{lk\}} = 0$. Then we can write the likelihood as:

$$L(w_1, \ldots, w_K) = \prod_{l=1}^{n} \prod_{k=1}^{K} P(Y^l = k | X^L = x; w)^{1_{\{lk\}}} = \prod_{l=1}^{n} \prod_{k=1}^{K} \left(\frac{\exp(w_k \cdot x^l)}{\sum_r \exp(w_r \cdot x^l)} \right)^{1_{\{lk\}}}.$$

Taking ln**:**

$$\ell(w_1, \ldots, w_K) = \sum_{l=1}^{n} \sum_{k=1}^{K} 1_{\{lk\}} \left(w_k \cdot x^l - \ln \sum_r \exp(w_r \cdot x^l) \right).$$

Adding the L_2 regularization term:

$$\ell(w_1, \ldots, w_K) = \sum_{l=1}^{n} \sum_{k=1}^{K} 1_{\{lk\}} \left(w_k \cdot x^l - \ln \sum_r \exp(w_r \cdot x^l) \right) - \frac{\lambda}{2} \sum_{k=1}^{K} \|w_k\|^2.$$

b. Taking derivative with respect to w_i:

$$\begin{aligned}\frac{\partial}{\partial w_i}\ell(w_1,\ldots,w_K) &= \sum_{t=1}^{n}\left(1_{\{li\}}\,x^l - \frac{\exp(w_i\cdot x^l)\,x^l}{\sum_r \exp(w_r\cdot x^l)}\right) - \lambda w_i \\ &= \sum_{t=1}^{n}\left(1_{\{li\}} - P(Y^l = i|X^l)\right)x^l - \lambda w_i.\end{aligned}$$

c. Then the update rule with gradient ascent for w_i is:

$$w_i \leftarrow w_i + \nu\sum_{t=1}^{n}\left(1_{\{li\}} - P(Y^l = i|X^l)\right)x^l - \nu\lambda w_i.$$

This will converge to a global maximum since it is a concave function.

Linear Regression and Logistic Regression:

definitions [revisited], and a common property

CMU, 2004 fall, T. Mitchell, Z. Bar-Joseph, HW2, pr. 4

Linear Regression and Logistic Regression: Definitions [revisited]

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,

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

b. Show that for each of the two regression models above, $\hat{\beta}$ 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 $\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} E[Y|X = x_i, \beta = \hat{\beta}]x_i$**.**