Accompanying Notes for *Of Neural Networks and Kernel Machines* Day 1: Neural Networks

Marco Del Vecchio

19/07/2017

1 Linear Classification Models

1.1 Logistic regression as an example of a probabilistic generative model

Let us consider a binary classification problem with training data given by

$$\{(\boldsymbol{x}^{(n)}, y^{(n)}) \in \mathbb{R}^D \times \{1, 2\} \mid n = 1, \ldots, N\},$$

Our goal is to classify an observation $\boldsymbol{x}^{(n)}$ as belonging to either class 1, $\mathcal{C}_1$, or class 2, $\mathcal{C}_2$. One way of achieving this is to use a linear discriminant function for logistic regression.

Definition 1 (Linear Discriminant Function for Logistic Regression)**.**

$$h(\boldsymbol{x}; \boldsymbol{w}) = \sigma(\boldsymbol{w}^T \boldsymbol{x} + w_0), \quad f(\boldsymbol{x}) = \begin{cases} 1 & \text{if} \quad h(\boldsymbol{x}) \geq \frac{1}{2} \\ 2 & \text{if} \quad h(\boldsymbol{x}) < \frac{1}{2} \end{cases}$$

where $\sigma(a) = \frac{1}{1+e^{-a}}$ *is the so called logistic or sigmoid function.*

Thanks to the fact that we have used the logistic function to squash the output of $\boldsymbol{w}^T \boldsymbol{x}$ into the interval $[0, 1]$, $h(\boldsymbol{x}; \boldsymbol{w})$ is now meaningful as a probability statement. Hence, we can set

$$\mathbb{P}(\mathcal{C}_1|\boldsymbol{x}) = h(\boldsymbol{x}; \boldsymbol{w}) = 1 - \mathbb{P}(\mathcal{C}_2|\boldsymbol{x}).$$

If we were to stop here, we would have an example of a probabilistic *discriminative* model because we are only modelling $\mathbb{P}(\mathcal{C}_k|\boldsymbol{x})$, $k \in \{1, 2\}$. However, logistic regression is *also* an example of a probabilistic *generative* model, that is of a model where the quantities $\mathbb{P}(\mathcal{C}_k)$, and $\mathbb{P}(\boldsymbol{x}|\mathcal{C}_k)$, $k \in \{1, 2\}$ are modelled.

Remark 1. *Suppose that*

$$\mathbb{P}(\mathcal{C}_1) = p = 1 - \mathbb{P}(\mathcal{C}_2),$$
$$\mathbb{P}(\boldsymbol{x}|\mathcal{C}_k) = \exp(A(\boldsymbol{\theta}_k) + B(\boldsymbol{x}, \boldsymbol{\phi}) + \boldsymbol{\theta}_k^T \boldsymbol{x}),$$

$k \in \{1, 2\}$*. That is, assume that the class-conditional densities are members of the exponential family of distributions, where the parameters* $\boldsymbol{\theta}_k$ *and* $\boldsymbol{\phi}$ *control the shape of the distribution.*

Then $\mathbb{P}(\mathcal{C}_1|\boldsymbol{x}) = \sigma(\boldsymbol{w}^T \boldsymbol{x} + w_0)$ *exactly, for a suitable choice of parameters* $\boldsymbol{w} \in \mathbb{R}^D$ *and bias* w_0*, where* $\sigma : \mathbb{R} \to \mathbb{R}$ *is the logistic function.*

Proof. By Bayes' Theorem,

$$\mathbb{P}(\mathcal{C}_1|\boldsymbol{x}) = \frac{\mathbb{P}(\boldsymbol{x}|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1)}{\mathbb{P}(\boldsymbol{x}|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1) + \mathbb{P}(\boldsymbol{x}|\mathcal{C}_2)\mathbb{P}(\mathcal{C}_2)}.$$

Dividing both numerator and denominator by $\mathbb{P}(\boldsymbol{x}|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1)$ we obtain

$$\begin{aligned}\mathbb{P}(\mathcal{C}_1|\boldsymbol{x}) &= \frac{1}{1+\frac{\mathbb{P}(\boldsymbol{x}|\mathcal{C}_2)\mathbb{P}(\mathcal{C}_2)}{\mathbb{P}(\boldsymbol{x}|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1)}} \\ &= \frac{1}{1+\left(\frac{\mathbb{P}(\boldsymbol{x}|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1)}{\mathbb{P}(\boldsymbol{x}|\mathcal{C}_2)\mathbb{P}(\mathcal{C}_2)}\right)^{-1}} \\ &= \frac{1}{1+\exp\left(-\log\left(\frac{\mathbb{P}(\boldsymbol{x}|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1)}{\mathbb{P}(\boldsymbol{x}|\mathcal{C}_2)\mathbb{P}(\mathcal{C}_2)}\right)\right)} \\ &= \frac{1}{1+e^{-a}}\end{aligned}$$

where $a = \log\left(\frac{\mathbb{P}(\boldsymbol{x}|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1)}{\mathbb{P}(\boldsymbol{x}|\mathcal{C}_2)\mathbb{P}(\mathcal{C}_2)}\right)$.
Hence, it remains to show that $a = \boldsymbol{w}^T\boldsymbol{x} + w_0$ for some $\boldsymbol{w} \in \mathbb{R}^D$ and $w_0 \in \mathbb{R}$:

$$\begin{aligned}a &= \log\left(\frac{\mathbb{P}(\boldsymbol{x}|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1)}{\mathbb{P}(\boldsymbol{x}|\mathcal{C}_2)\mathbb{P}(\mathcal{C}_2)}\right) \\ &= \log(\mathbb{P}(\boldsymbol{x}|\mathcal{C}_1)\mathbb{P}(\mathcal{C}_1)) - \log(\mathbb{P}(\boldsymbol{x}|\mathcal{C}_2)\mathbb{P}(\mathcal{C}_2)) \\ &= \log\mathbb{P}(\boldsymbol{x}|\mathcal{C}_1) + \log\mathbb{P}(\mathcal{C}_1) - \log\mathbb{P}(\boldsymbol{x}|\mathcal{C}_2) - \log\mathbb{P}(\mathcal{C}_2)\end{aligned}$$

having assumed that $\mathbb{P}(\mathcal{C}_1) = p = 1-\mathbb{P}(\mathcal{C}_2)$, and that $\mathbb{P}(\boldsymbol{x}|\mathcal{C}_k) = \exp(A(\boldsymbol{\theta}_k)+B(\boldsymbol{x},\boldsymbol{\phi})+\boldsymbol{\theta}_k^T\boldsymbol{x})$, this yields

$$\begin{aligned}a &= A(\boldsymbol{\theta}_1) + B(\boldsymbol{x},\boldsymbol{\phi}) + \boldsymbol{\theta}_1^T\boldsymbol{x} + \log p - A(\boldsymbol{\theta}_2) - B(\boldsymbol{x},\boldsymbol{\phi}) - \boldsymbol{\theta}_2^T\boldsymbol{x} - \log(1-p) \\ &= A(\boldsymbol{\theta}_1) - A(\boldsymbol{\theta}_2) + (\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^T\boldsymbol{x} + \log\left(\frac{p}{1-p}\right).\end{aligned}$$

That is,

$$a = \boldsymbol{w}^T\boldsymbol{x} + w_0,$$

with

$$\begin{aligned}\boldsymbol{w} &= \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \\ w_0 &= A(\boldsymbol{\theta}_1) - A(\boldsymbol{\theta}_2) + \log\left(\frac{p}{1-p}\right) = A(\boldsymbol{\theta}_1) - A(\boldsymbol{\theta}_2) + \log\left(\frac{\mathbb{P}(\mathcal{C}_1)}{\mathbb{P}(\mathcal{C}_2)}\right)\end{aligned}$$

□

2 Training

2.1 Choice of learning rate

Theorem 1 (Taylor Theorem)**.** *Let $k \geq 1$ be an integer and let the function $f : \mathbb{R} \to \mathbb{R}$ be k times differentiable at the point $c \in \mathbb{R}$ then, there exist a function $R_k : \mathbb{R} \to \mathbb{R}$ such that*

$$f(x) = \sum_{i=0}^{k} \frac{f^{(i)(c)}}{i!}(x-c)^i + R_k(x)(x-a)^k,$$

and $\lim_{x\to c} R_k = 0$.

Let us focus on a single arbitrary weight $w \in \mathbb{R}$, with $G = G(w)$, minimised at

$$w^* = \arg\min_w G(w).$$

A Taylor expansion of $G(w)$ around the value of the weight at time t, $w^{(t)}$ yields

$$G(w) = G(w^{(t)}) + (w - w^{(t)})G'(w^{(t)}) + \frac{1}{2}(w - w^{(t)})^2 G''(w^{(t)}) + R_2(w)(w - w^{(t)})^2$$

Replacing $G(w)$ with its quadratic approximation

$$G(w) \sim G(w^{(t)}) + (w - w^{(t)})G'(w^{(t)}) + \frac{1}{2}(w - w^{(t)})^2 G''(w^{(t)}),$$

we find that $G(w)$ is minimised at the point w^* satisfying

$$G'(w^*) = 0 \implies G'(w^{(t)}) + (w^* - w^{(t)})G''(w^{(t)}) = 0.$$

Rearranging we find that

$$w^* = w^{(t)} - \frac{1}{G''(w^{(t)})}G'(w^{(t)})$$

Now, if we compare the expression above with the weight update step in stochastic gradient descent

$$w^{(t+1)} = w^{(t)} - \eta \nabla G(w^{(t)})$$

we can see that at the optimum, η should be equal to $\frac{1}{G''(w^{(t)})}$ that is, η should be equal to the inverse of the curvature of the error function at the current value of the weight.
More generally, for a vector of weights $\boldsymbol{w}$ the same argument leads the conclusion that the optimal choice of learning rate η is

$$\eta = \boldsymbol{H}^{-1}(\boldsymbol{w}^{(n)})$$

where $\boldsymbol{H}(\boldsymbol{w})$ is the Hessian matrix of second derivatives:

$$\boldsymbol{H}^{-1}(\boldsymbol{w}) = \left(\frac{\partial^2 G(\boldsymbol{w})}{\partial w_i \partial w_j} \right)_{ij}.$$

This suggests that the optimal update for a component w_i depends on the slope and curvature of G in every direction, not just that of w_i.