Introduction to Machine Learning

Answers to Exercise 5

Probabilistic Modelling & Decision Theory

Jingtao Min

July 19, 2022

1 Multiclass logistic regression

Posterior probabilities for multiclass logistic regression can be given as a softmax transformation of hyperplanes:

$$f_k(y, x, a_1, \cdots a_K) = P(y = k|X = \mathbf{x}, a_1, \cdots a_K) = \frac{\exp\left(\mathbf{a}_k^T \mathbf{x}\right)}{\sum_j \exp\left(\mathbf{a}_j^T \mathbf{x}\right)} = \frac{\exp\left(s_k\right)}{\sum_j \exp\left(s_j\right)} \tag{1}$$

Consider the derivative $D_{ki} = \frac{\partial f_k}{\partial s_i}$:

(a) The derivative element D_{ki} *per se* is a scalar, but the derivatives obtained in this form constitute a $K \times K$ matrix.

(b) The derivative is expressed as:

$$D_{ki} = \frac{\partial}{\partial s_i}\left(\frac{\exp\left(s_k\right)}{\sum_j \exp\left(s_j\right)}\right) = \frac{\delta_{ik} e^{s_k} \sum_j e^{s_j} - e^{s_k + s_i}}{\left(\sum_j e^{s_j}\right)^2} = \delta_{ik} f_k - f_k f_i = \left(\delta_{ik} - f_i\right) f_k \tag{2}$$

(c) In practice the derivative by $\mathbf{a}_i$ is sought, as these vectors are the unknowns where training takes place. Given a class k and the derivative takes the form

$$\frac{\partial f_k}{\partial \mathbf{a}_i} = \frac{\partial f_k}{\partial s_i} \frac{\partial s_i}{\partial \mathbf{a}_i} = D_{ki} \mathbf{x} \tag{3}$$

(d) Softmax implicit requires that s_i cannot be arbitrarily large. For instance, if $s_i > 10^3$, explicit computation of $\exp\left(s_i\right)$ is not possible due to overflow, as it reaches limit of the double-precision exponent. It might, however, be possible if proper scaling is used beforehand, as follows.

(e) First, pick the extreme of $|s_j|$, i.e. $s_{i^*} = \min_{i \in K} |s_i|$; then, substract this value from other weights:

$$f_k = \frac{\exp(s_k)}{\sum_i \exp(s_i)} = \frac{\exp(s_k - s_{i^*})}{\sum_i \exp(s_i - s_{i^*})} = \frac{\exp(s'_k)}{1 + \sum_{i \neq i^*} \exp(s'_i)} \tag{4}$$

2 Decision theory

Consider an binary option investment problem, where the information is encoded as a vector $\mathbf{x} \in \mathbb{R}^d$.

(a) Note that binary option means all or nothing. Therefore, the probability of gaining profit from a binary option is always a binary/Bernoulli distribution. Denoting the weights used in the logistic model as $\mathbf{w}$, the estimated conditional probability of gaining profit from the binary option is given by:

$$P(y|\mathbf{x}, \mathbf{w}) = \text{Ber}\left(y; \sigma(\mathbf{w}^T \mathbf{x})\right) \tag{5}$$

where $\text{Ber}\left(y; \sigma(\mathbf{w}^T \mathbf{x})\right)$ is Bernoulli distribution with $p = \sigma(\mathbf{w}^T \mathbf{x})$. $y = 1$ means the investment profits, while $y = 0$ means no payoff.

The action set $\mathcal{A}$ includes two actions: invest in {secure investment, binary option}, denoted by $a = 0$ and $a = 1$, respectively. Finally, under these definitions of events $\mathcal{Y}$ and actions $\mathcal{A}$ we come to the definition of the cost function as follows Tab. 1. The values of the costs are computed via how much money is lost compared to the best action scenario under given event y. I didn't think of assigning costs like this *a priori*, but followed what the reference solution seems to imply.

Table 1: Cost function setup

Outcomes / Actions	$a = 0$ (Secure)	$a = 1$ (Binary)
$y = 0$ (No binary payoff)	0	1200
$y = 1$ (Binary profit)	400	0

(b) If the investor decides to buy a binary option, the expected cost:

$$\mathbb{E}_y[C(y|a=1)|\mathbf{x}] = 1200\,(1 - p(\mathbf{x})) = 1200\left(1 - \sigma(\mathbf{w}^T\mathbf{x})\right) \tag{6}$$

(c) The decision rule derives from maximum expected utility / minimum expected cost. Given $\mathbf{x}$ the expected cost for binary option is already derived. The expected cost for secure investment is given by:

$$\mathbb{E}_y[C(y|a=0)|\mathbf{x}] = 400\,p(\mathbf{x}) = 400\,\sigma(\mathbf{w}^T\mathbf{x}) \tag{7}$$

We have the inequality:

$$\mathbb{E}_y[C(y|a=0)|\mathbf{x}] \leq \mathbb{E}_y[C(y|a=1)|\mathbf{x}] \quad \left(p(\mathbf{x}) \leq \frac{3}{4} = 0.75\right) \tag{8}$$

Hence the decision rule:

$$f(\mathbf{x}) : \begin{cases} a = 0, & p(\mathbf{x}) \leq 0.75 \\ a = 1, & p(\mathbf{x}) > 0.75 \end{cases} \tag{9}$$

(d) We denote the binary outcome of the model as random variable $\widetilde{\mathcal{Y}}$ and follow the same convention as y. If the model outputs binary outcome is only correct with probability p, we can expand the actions into four (2×2) combinations of model output and actual decisions:

Table 2: Cost function setup

Outcomes / Actions	$\widetilde{y} = 0$ (No profit) $a = 0$ (Secure)	$\widetilde{y} = 0$ (No profit) $a = 1$ (Binary)	$\widetilde{y} = 1$ (Profit) $a = 0$ (Secure)	$\widetilde{y} = 1$ (Profit) $a = 1$ (Binary)
$y = 0$ (No profit)	0	1200	0	1200
$y = 1$ (Profit)	400	0	400	0

The expected cost for each decision:

$$\begin{aligned} &\mathbb{E}_y[C(y|\widetilde{y}=0, a=0)] = 400\,(1-p) \\ &\mathbb{E}_y[C(y|\widetilde{y}=0, a=1)] = 1200\,p \\ &\mathbb{E}_y[C(y|\widetilde{y}=1, a=0)] = 400\,p \\ &\mathbb{E}_y[C(y|\widetilde{y}=1, a=1)] = 1200\,(1-p) \end{aligned} \tag{10}$$

And finally the decision function:

$$f(\widetilde{y}=0) : \begin{cases} a = 0, & (p \geq 0.25) \\ a = 1, & (p < 0.25) \end{cases} \qquad f(\widetilde{y}=1) : \begin{cases} a = 0, & (p \leq 0.75) \\ a = 1, & (p > 0.75) \end{cases} \tag{11}$$

3 Naive Bayes estimate

Consider binary classification problem, where $\mathcal{Y}$ is the set of labels and $\mathcal{X} = \mathbb{N}^d$ is a d-dimensional feature space (each element is a natural number). A training set $D = \{(\mathbf{x}_i, y_i)\}$ of n samples is given, where all features are geometric distributed with parameters $\hat{p}_j$ $(j \in \{1, 2, \cdots d\})$.

(a) Let $\{z_i\}_{i=1}^m$ be m i.i.d. observations of a p-geometric distributed random variable. The likelihood function given parameter p:

$$\begin{aligned} P\left(Z=k|p\right) &= (1-p)^{k-1}p \\ P\left(\{z_i\}_{i=1}^m|p\right) &= \prod_{i=1}^m (1-p)^{z_i-1}p = (1-p)^{\sum_{i=1}^m z_i - m}\, p^m \\ \ln P\left(\{z_i\}_{i=1}^m|p\right) &= \left(\sum_{i=1}^m z_i - m\right)\ln(1-p) + m\ln p \end{aligned} \tag{12}$$

Maximizing the likelihood is equivalently maximizing the log likelihood, and the optimum satisfies:

$$\frac{\partial}{\partial p}\ln P\left(\{z_i\}_{i=1}^m|p\right) = \frac{m}{p} - \frac{\sum_{i=1}^m z_i - m}{1-p} = 0 \quad \Longrightarrow \quad p = \frac{m}{\sum_{i=1}^m z_i} = \frac{1}{\bar{z_i}} \tag{13}$$

(b) The likelihood function given the training dataset:

$$\begin{aligned} P(\mathbf{X}|Y) &= \sum_{i=1}^d P(X_i|Y) = \prod_{j=1}^d \left(1-\hat{p}_{j,y}\right)^{x_j-1}\hat{p}_{j,y} \\ P(\mathbf{x}_{i_{i=1}}^{n_y}|Y=y) &= \prod_{j=1}^d \left(1-\hat{p}_{j,y}\right)^{\sum_{i=1}^{n_y} x_j^{(i)} - n_0}\hat{p}_{j,y}^{n_y} \end{aligned} \tag{14}$$

According to the previous conclusion, we have the estimates

$$\hat{p}_{j,y} = \frac{\{\text{Count } y_i = y\}}{\sum_{y_i=y} x_j^{(i)}} \tag{15}$$

Once the parameters are estimated via MLE, the joint distribution can be stated:

$$P(\mathbf{X},Y) = P(\mathbf{X}|Y)P(Y) = P(Y)\sum_{i=1}^d P(X_i|Y) = p_y\prod_{j=1}^d \left(1-\hat{p}_{j,y}\right)^{x_j-1}\hat{p}_{j,y} \tag{16}$$

(c) Given a new data $\mathbf{x}$, the prediction is computed via the posterior distribution:

$$\begin{aligned} P(Y|\mathbf{X}=\mathbf{x}) &\propto p_y\prod_{j=1}^d \left(1-\hat{p}_{j,y}\right)^{x_j-1}\hat{p}_{j,y} \\ P(Y=0|\mathbf{X}=\mathbf{x}) &= Cp_0\prod_{j=1}^d \left(1-\hat{p}_{j,0}\right)^{x_j-1}\hat{p}_{j,0} \\ P(Y=1|\mathbf{X}=\mathbf{x}) &= Cp_1\prod_{j=1}^d \left(1-\hat{p}_{j,1}\right)^{x_j-1}\hat{p}_{j,1} \end{aligned} \tag{17}$$

The two probabilities can be compared by taking their quotient:

$$\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})} = \frac{p_1}{p_0}\prod_{j=1}^d \left(\frac{1-\hat{p}_{j,1}}{1-\hat{p}_{j,i}}\right)^{x_j-1}\left(\frac{\hat{p}_{j,1}}{\hat{p}_{j,0}}\right) = \frac{C_1}{C_0}\prod_{j=1}^d \left(\frac{1-\hat{p}_{j,1}}{1-\hat{p}_{j,i}}\right)^{x_j-1} \tag{18}$$

The prediction boundary can be given by setting $P(Y=1|\mathbf{X}=\mathbf{x}) = P(Y=0|\mathbf{X}=\mathbf{x})$, equivalently:

$$\begin{aligned} \prod_{j=1}^d \left(\frac{1-\hat{p}_{j,1}}{1-\hat{p}_{j,i}}\right)^{x_j-1} &= \frac{p_0}{p_1}\prod_{j=1}^d \frac{\hat{p}_{j,0}}{\hat{p}_{j,1}} \\ \sum_{j=1}^d (x_j-1)\ln\left(\frac{1-\hat{p}_{j,1}}{1-\hat{p}_{j,0}}\right) = \ln\prod_{j=1}^d \left(\frac{1-\hat{p}_{j,1}}{1-\hat{p}_{j,i}}\right)^{x_j-1} &= \ln\frac{p_0}{p_1} + \sum_{j=1}^d \ln\frac{\hat{p}_{j,0}}{\hat{p}_{j,1}} \equiv a \\ a_j = \ln\left(\frac{1-\hat{p}_{j,1}}{1-\hat{p}_{j,0}}\right) \quad \Longrightarrow \quad \sum_{j=1}^d a_j(x_j-1) &= a \\ \Longrightarrow \quad \mathbf{a}^T\mathbf{x} = \sum_{j=1}^d a_j x_j = a + \sum_{j=1}^d a_j &\equiv b \end{aligned} \tag{19}$$

Therefore the boundary is a hyperplane.

(d) Assuming $x_2 \equiv x_3 \equiv \cdots x_d$ are indentical features, we can assert that $\hat{p}_{2,y} \equiv \hat{p}_{3,y} \equiv \cdots \hat{p}_{d,y}$, the coefficients for the hyperplane:

$$a_{2\cdots d} = \ln \frac{1-\hat{p}_{2,1}}{1-\hat{p}_{2,0}}, \quad a = \ln \frac{p_0}{p_1} + \ln \frac{\hat{p}_{1,0}}{\hat{p}_{1,1}} + (d-1) \ln \frac{\hat{p}_{2,0}}{\hat{p}_{2,1}} = \ln \frac{p_0 \hat{p}_{1,0} \hat{p}_{2,0}^{d-1}}{p_1 \hat{p}_{1,1} \hat{p}_{2,1}^{d-1}} \tag{20}$$

And the hyperplane using the naive approach is given by:

$$\left(\ln \frac{1-\hat{p}_{1,1}}{1-\hat{p}_{1,0}} \right) x_1 + (d-1) \left(\ln \frac{1-\hat{p}_{2,1}}{1-\hat{p}_{2,0}} \right) x_2 = \ln \frac{p_0}{p_1} \frac{\hat{p}_{1,0}(1-\hat{p}_{1,1})}{\hat{p}_{1,1}(1-\hat{p}_{1,0})} \frac{\hat{p}_{2,0}^{d-1}(1-\hat{p}_{2,1})^{d-1}}{\hat{p}_{2,1}^{d-1}(1-\hat{p}_{2,0})^{d-1}} \tag{21}$$

It is however clear that only two features are present in the data. Therefore the appropriate hyperplane that maximizes the posterior is:

$$\left(\ln \frac{1-\hat{p}_{1,1}}{1-\hat{p}_{1,0}} \right) x_1 + \left(\ln \frac{1-\hat{p}_{2,1}}{1-\hat{p}_{2,0}} \right) x_2 = \ln \frac{p_0}{p_1} \frac{\hat{p}_{1,0}(1-\hat{p}_{1,1})}{\hat{p}_{1,1}(1-\hat{p}_{1,0})} \frac{\hat{p}_{2,0}(1-\hat{p}_{2,1})}{\hat{p}_{2,1}(1-\hat{p}_{2,0})} \tag{22}$$

4 Bias-variance trade-off

Consider a dataset of n i.i.d. samples $\{x_i, y_i\}_{i=1}^d$ and least squares regression, the prediction error can be decomposed into (bias^2) + (variance) + (noise):

$$\mathbb{E}_{(x,y),D} \left[\left(y - \hat{f}_D(x) \right)^2 \right] = \mathbb{E}_x \left[\left(f^*(x) - \bar{f}(x) \right)^2 \right] + \mathbb{E}_x \left[\mathrm{Var}_D \left(\hat{f}_D(x) \right) \right] + \mathbb{E}_{x,y} \left[\left(y - f^*(x) \right)^2 \right] \tag{23}$$

(a) If the bias increases but the variance decreases at the same time, it is possible to reduce prediction error;

(b) For an increasing amount of data, both the bias and the variance are expected to decrease;

(c) A strictly larger hypothesis class $\mathcal{H}_{\text{old}} \subset \mathcal{H}_{\text{new}}$, then the bias must be smaller or equal to the bias considering the old hypothesis class.

(d) If $n \to \infty$ and the finite hypothesis class includes the optimal model f^*, then the prediction error should only depend on the noise and the opitmal hypothesis.

5 Distribution shifts