Probabilistic classification - discriminative models

Course of Machine Learning
Master Degree in Computer Science
University of Rome "Tor Vergata"
a.a. 2024-2025

Giorgio Gambosi

Generalized linear models

In the cases considered above, the posterior class distributions $p(C_k|\mathbf{x})$ are sigmoidal or softmax with argument given by a linear combination of features in $\mathbf{x}$, i.e., they are a instances of **generalized linear models**

A **generalized linear model** (GLM) is a function

$$h(\mathbf{x}) = f(\mathbf{w}^T\mathbf{x} + b) = f(\overline{\mathbf{w}}^T\overline{\mathbf{x}})$$

where f (usually called the *response function*) is in general a non linear function.

Each iso-surface of $h(\mathbf{x})$, such that by definition $h(\mathbf{x}) = c$ (for some constant c), is such that

$$f(\overline{\mathbf{w}}^T\overline{\mathbf{x}}) = c$$

and

$$\overline{\mathbf{w}}^T\overline{\mathbf{x}} = f^{-1}(t) = c'$$

(c' constant).

Hence, iso-surfaces of a GLM are hyper-planes, thus implying that boundaries are hyperplanes themselves.

Exponential families and GLM

Let us assume we wish to predict a random variable t as a function of a different set of random variables $\mathbf{x}$. By definition, a prediction model for this task is a GLM if the following hypotheses hold:

1. the conditional distribution $p(t|\mathbf{x})$ belongs to the exponential family: that is, we may write it as

$$p(t|\mathbf{x}) = \frac{1}{s}g(\boldsymbol{\theta}(\mathbf{x}))f\left(\frac{t}{s}\right)e^{\frac{1}{s}\boldsymbol{\theta}(\mathbf{x})^T\mathbf{u}(t)}$$

   for suitable $g, \boldsymbol{\theta}, \mathbf{u}$

2. for any $\mathbf{x}$, we wish to predict the expected value of $\mathbf{u}(t)$ given $\mathbf{x}$, that is $E[\mathbf{u}(t)|\mathbf{x}]$

3. $\boldsymbol{\theta}(\mathbf{x})$ (the **natural parameter**) is a linear combination of the features, $\boldsymbol{\theta}(\mathbf{x}) = \overline{\mathbf{w}}^T\overline{\mathbf{x}}$

GLM and normal distribution

1. Assume $t \in \mathbb{R}$, and $p(t|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu(\mathbf{x}))^2}{2\sigma^2}}$ is a normal distribution with mean $\mu(\mathbf{x})$ and constant variance σ^2: it is easy to verify that

$$\boldsymbol{\theta}(\mathbf{x}) = \begin{pmatrix} \theta_1(\mathbf{x}) \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \mu(\mathbf{x})/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix}$$

and $\mathbf{u}(t) = t$

2. we wish to predict the value of $E[\mathbf{u}(t)|\mathbf{x}] = E[t|\mathbf{x}] = \mu(\mathbf{x})$ as $h(\mathbf{x})$, then

$$h(\mathbf{x}) = \mu(\mathbf{x}) = \sigma^2\theta_1(\mathbf{x})$$

3. we assume $\theta_1(\mathbf{x})$ is a linear combination of the features $\theta_1(\mathbf{x}) = \overline{\mathbf{w}}^T\overline{\mathbf{x}}$

Then,

$$h(\mathbf{x}) = \sigma^2\overline{\mathbf{w}}^T\overline{\mathbf{x}}$$

and a linear regression $h(\mathbf{x}) = \overline{\mathbf{u}}^T\overline{\mathbf{x}}$ results with $u_i = \sigma^2 w_i, i = 0, \ldots, d$.

GLM and Bernoulli distribution

1. Assume $t \in \{0, 1\}$, and $p(t|\mathbf{x}) = \pi(\mathbf{x})^t(1-\pi(\mathbf{x}))^{1-t}$ is a Bernoulli distribution with parameter $\pi(\mathbf{x})$: then, the natural parameter $\theta(\mathbf{x})$ can be shown to be

$$\theta(\mathbf{x}) = \log\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}$$

and $\mathbf{u}(t) = t$

2. we wish to predict the value of $E[\mathbf{u}(t)|\mathbf{x}] = E[t|\mathbf{x}] = p(t=1|\mathbf{x}) = \pi(\mathbf{x})$ as $h(\mathbf{x})$, then

$$h(\mathbf{x}) = \frac{1}{1+e^{-\theta(\mathbf{x})}}$$

3. we assume $\theta(\mathbf{x})$ is a linear combination of the features $\theta(\mathbf{x}) = \overline{\mathbf{w}}^T\overline{\mathbf{x}}$

Then, a logistic regression derives

$$h(\mathbf{x}) = \frac{1}{1+e^{-\overline{\mathbf{w}}^T\overline{\mathbf{x}}}}$$

GLM and categorical distribution

1. Assume $t \in \{1, \ldots, K\}$, and $p(t|\mathbf{x}) = \prod_{i=1}^K \pi_i(\mathbf{x})^{t_i}$ (where $t_i = 1$ if $t = i$ and $t_i = 0$ otherwise) is a categorical distribution with probabilities $\pi_1(\mathbf{x}), \ldots, \pi_K(\mathbf{x})$: the natural parameter is then $\boldsymbol{\theta}(\mathbf{x}) = (\theta_1(\mathbf{x}), \ldots, \theta_K(\mathbf{x}))^T$, with

$$\theta_i(\mathbf{x}) = \log\frac{\pi_i(\mathbf{x})}{\pi_K(\mathbf{x})} = \log\frac{\pi_i(\mathbf{x})}{1-\sum_{j=1}^{K-1}\pi_j(\mathbf{x})}$$

and $\mathbf{u}(t) = (t_1, \ldots, t_K)^T$ is the 1-to-K representation of t

2. we wish to predict the expectations $E[u_i(t)|\mathbf{x}] = p(t = i|\mathbf{x})$ as

$$h_i(\mathbf{x}) = p(t = i|\mathbf{x}) = \pi_i(\mathbf{x}) = \pi_K(\mathbf{x})e^{\theta_i(\mathbf{x})}$$

   Since $\sum_{i=1}^{K} \pi_i(\mathbf{x}) = \pi_K(\mathbf{x}) \sum_{i=1}^{K} e^{\theta_i(\mathbf{x})} = 1$, it derives

$$\pi_K(\mathbf{x}) = \frac{1}{\sum_{i=1}^{K} e^{\theta_i(\mathbf{x})}} \qquad \text{and} \qquad \pi_i(\mathbf{x}) = \frac{e^{\theta_i(\mathbf{x})}}{\sum_{i=1}^{K} e^{\theta_i(\mathbf{x})}}$$

3. we assume all $\theta_i(\mathbf{x})$ are linear combinations of the features $\theta_i(\mathbf{x}) = \overline{\mathbf{w}}_i^T \overline{\mathbf{x}}$

Then, a softmax regression results, with

$$\begin{aligned} h_i(\mathbf{x}) &= \frac{e^{\overline{\mathbf{w}}_i^T \overline{\mathbf{x}}}}{\sum_{j=1}^{K} e^{\overline{\mathbf{w}}_j^T \overline{\mathbf{x}}}} \qquad i = 1, \ldots, K-1 \\ h_K(\mathbf{x}) &= \frac{1}{\sum_{j=1}^{K} e^{\overline{\mathbf{w}}_j^T \overline{\mathbf{x}}}} \end{aligned}$$

GLM and additional regressions

Other regression types can be defined by considering different models for $p(t|\mathbf{x})$. For example,

Poisson distribution

1. Assume $t \in \{0, \ldots, \}$ is a non negative integer (for example we are interested to count data), and $p(t|\mathbf{x}) = \frac{\lambda(\mathbf{x})^t}{y!} e^{-\lambda(\mathbf{x})}$ is a Poisson distribution with parameter $\lambda(\mathbf{x})$: then, the natural parameter $\theta(\mathbf{x})$ is

$$\theta(\mathbf{x}) = \log \lambda(\mathbf{x})$$

   and $\mathbf{u}(t) = t$

2. we wish to predict the expectation of $E[\mathbf{u}(t)|\mathbf{x}] = E[t|\mathbf{x}] = \lambda(\mathbf{x})$ as

$$h(\mathbf{x}) = \lambda(\mathbf{x}) = e^{\theta(\mathbf{x})}$$

3. we assume $\theta(\mathbf{x})$ is a linear combination of the features $\theta(\mathbf{x}) = \overline{\mathbf{w}}^T \overline{\mathbf{x}}$

Then, a Poisson regression derives

$$h(\mathbf{x}) = e^{\overline{\mathbf{w}}^T \overline{\mathbf{x}}}$$

Exponential distribution

1. Assume $t \in [0, \infty)$ is a non negative real (for example we are interested to time intervals), and $p(t|\mathbf{x}) = \lambda(\mathbf{x})e^{-\lambda(\mathbf{x})t}$ is an exponential distribution with parameter $\lambda(\mathbf{x})$: then, the natural parameter $\theta(\mathbf{x})$ is

$$\theta(\mathbf{x}) = -\lambda(\mathbf{x})$$

   and $\mathbf{u}(t) = t$

2. we wish to predict the value of $E[\mathbf{u}(t)|\mathbf{x}] = E[t|\mathbf{x}]$ as $h(\mathbf{x})$, then

$$h(\mathbf{x}) = \frac{1}{\lambda(\mathbf{x})} = -\frac{1}{\theta(\mathbf{x})}$$

3. we assume $\theta(\mathbf{x})$ is a linear combination of the features $\theta(\mathbf{x}) = \overline{\mathbf{w}}^T\overline{\mathbf{x}}$

Then, an exponential regression derives

$$h(\mathbf{x}) = -\frac{1}{\overline{\mathbf{w}}^T\overline{\mathbf{x}}}$$

Discriminative approach

In the discriminative approach we are interested in modeling $p(C_k|\mathbf{x})$: In particular, we may assume that such probability is a GLM and derive its coefficients (for example through ML estimation).

Comparison wrt the generative approach:

- Less information derived (we do not know $p(\mathbf{x}|C_k)$, thus we are not able to generate new data)
- Simpler method, usually a smaller set of parameters to be derived
- Better predictions, if the assumptions done with respect to $p(\mathbf{x}|C_k)$ are poor.

Logistic regression

Logistic regression is a GLM deriving from the hypothesis of a Bernoulli distribution of t, which results into

$$p(C_1|\mathbf{x}) = \sigma(\overline{\mathbf{w}}^T\overline{\mathbf{x}}) = \frac{1}{1 + e^{-\overline{\mathbf{w}}^T\overline{\mathbf{x}}}}$$

where, as always, base functions could also be applied.

The model is equivalent, for the binary classification case, to linear regression for the regression case.

Degrees of freedom

- Logistic regression requires $d + 1$ coefficients $b, w_1, \ldots, w_d$ to be derived from a training set
- A generative approach with gaussian distributions requires:
  - $2d$ coefficients for the means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$,
  - for each covariance matrix

$$\sum_{i=1}^{d} i = \frac{d(d+1)}{2} \quad \text{coefficients}$$

  - one prior class probability $p(C_1)$
- As a total, it results into $d(d + 1) + 2d + 1 = d(d + 3) + 1$ coefficients (if a unique covariance matrix is assumed $d(d + 1)/2 + 2d + 1 = d(d + 5)/2 + 1$ coefficients)

Maximum likelihood estimation

As stated above, we assume that targets of elements of the training set can be conditionally (with respect to model coefficients) modeled through a Bernoulli distribution. That is, assume

$$p(t_i|\mathbf{x}_i;\mathbf{w}) = p_i^{t_i}(1-p_i)^{1-t_i}$$

where $p_i = p(C_1|\mathbf{x}_i) = \sigma(a_i)$ and $a_i = \overline{\mathbf{w}}^T\overline{\mathbf{x}}_i$

Then, the likelihood of the training set targets $\mathbf{t}$ given $\mathbf{X}$ is

$$p(\mathbf{t}|\mathrm{X};\overline{\mathbf{w}}) = L(\overline{\mathbf{w}}|\mathrm{X},\mathbf{t}) = \prod_{i=1}^{n} p(t_i|\mathbf{x}_i;\overline{\mathbf{w}}) = \prod_{i=1}^{n} p_i^{t_i}(1-p_i)^{1-t_i}$$

and the log-likelihood is

$$l(\overline{\mathbf{w}}|\mathrm{X},\mathbf{t}) = \log L(\overline{\mathbf{w}}|\mathrm{X},\mathbf{t}) = \sum_{i=1}^{n}(t_i \log p_i + (1-t_i)\log(1-p_i))$$

- Since

$$\frac{\partial l}{\partial w_j} = \sum_{i=1}^{n} \frac{\partial \log p(\overline{\mathbf{w}}|\mathbf{x}_i,t_i)}{\partial p_i}\frac{\partial p_i}{\partial a_i}\frac{\partial a_i}{\partial w_j}$$

$$\frac{\partial \log p(\overline{\mathbf{w}}|\mathbf{x}_i,t_i)}{\partial p_i} = \frac{t_i}{p_i} - \frac{1-t_i}{1-p_i} = \frac{t_i(1-p_i)-p_i(1-t_i)}{p_i(1-p_i)} = \frac{t_i-p_i}{p_i(1-p_i)}$$

$$\frac{\partial p_i}{\partial a_i} = \frac{\partial \sigma(a_i)}{\partial a_i} = \sigma(a_i)(1-\sigma(a_i)) = p_i(1-p_i)$$

$$\frac{\partial a_i}{\partial w_j} = x_{ij}$$

and

$$\frac{\partial l}{\partial b} = \sum_{i=1}^{n} \frac{\partial \log p(\overline{\mathbf{w}}|\mathbf{x}_i,t_i)}{\partial p_i}\frac{\partial p_i}{\partial a_i}\frac{\partial a_i}{\partial b}$$

$$\frac{\partial a_i}{\partial b} = 1$$

- It results

$$\frac{\partial}{\partial w_j} l(\overline{\mathbf{w}}|\mathrm{X},\mathbf{t}) = \sum_{i=1}^{n}\frac{t_i-p_i}{p_i(1-p_i)}p_i(1-p_i)x_{ij}$$
$$= \sum_{i=1}^{n}(t_i-p_i)x_{ij} = \sum_{i=1}^{n}(t_i-\sigma(\overline{\mathbf{w}}^T\overline{\mathbf{x}}_i))x_{ij}$$

and

$$\frac{\partial}{\partial b} l(\overline{\mathbf{w}}|\mathrm{X},\mathbf{t}) = \sum_{i=1}^{n}(t_i-\sigma(\overline{\mathbf{w}}^T\overline{\mathbf{x}}_i))$$

- In vector notation

$$\nabla_{\overline{\mathbf{w}}} l(\overline{\mathbf{w}}|\mathrm{X}, \mathbf{t}) = \sum_{i=1}^{n} (t_i - \sigma(\overline{\mathbf{w}}^T \overline{\mathbf{x}}_i))\overline{\mathbf{x}}_i$$

To maximize the likelihood, we could apply a gradient ascent algorithm, where at each iteration the following update of the currently estimated **w** is performed

$$\begin{aligned}
\overline{\mathbf{w}}^{(j+1)} &= \overline{\mathbf{w}}^{(j)} + \alpha \nabla_{\overline{\mathbf{w}}} l(\overline{\mathbf{w}}|\mathrm{X}, \mathbf{t})|_{\overline{\mathbf{w}}^{(j)}} \\
&= \overline{\mathbf{w}}^{(j)} + \alpha \sum_{i=1}^{n} (t_i - \sigma((\overline{\mathbf{w}}^{(j)})^T \overline{\mathbf{x}}_i))\overline{\mathbf{x}}_i \\
&= \mathbf{w}^{(j)} + \alpha \sum_{i=1}^{n} (t_i - h^{(j)}(\mathbf{x}_i))\mathbf{x}_i
\end{aligned}$$

Logistic regression and GDA

- Observe that assuming $p(\mathbf{x}|C_1)$ are $p(\mathbf{x}|C_2)$ as multivariate normal distributions with same covariance matrix Σ results into a logistic $p(C_1|\mathbf{x})$.
- The opposite, however, is not true in general: in fact, GDA relies on stronger assumptions than logistic regression.
- The more the normality hypothesis of class conditional distributions with same covariance is verified, the more GDA will tend to provide the best models for $p(C_1|\mathbf{x})$
- Logistic regression relies on weaker assumptions than GDA: it is then less sensible from a limited correctness of such assumptions, thus resulting in a more robust technique
- Since $p(C_i|\mathbf{x})$ is logistic under a wide set of hypotheses about $p(\mathbf{x}|C_i)$, it will usually provide better solutions (models) in all such cases, while GDA will provide poorer models as far as the normality hypotheses is less verified.

Softmax regression

In order to extend the logistic regression approach to the case $K > 2$, let us consider the matrix $\overline{W} = (\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_K)$ of model coefficients, of size $(d+1) \times K$, where $\mathbf{w}_j$ is the $d+1$-dimensional vector of coefficients for class C_j. In this case, the likelihood is defined as

$$p(\mathbf{T}|\mathrm{X}, \overline{\mathrm{W}}) = \prod_{i=1}^{n} \prod_{k=1}^{K} y_{ik}^{t_{ik}}$$

where

$$y_{ik} = p(C_k|\mathbf{x}_i) = \frac{e^{\overline{\mathbf{w}}_k^T \overline{\mathbf{x}}_i}}{\sum_{r=1}^{K} e^{\overline{\mathbf{w}}_r^T \overline{\mathbf{x}}_i}}$$

and $\mathbf{T}$ is the $n \times K$ matrix where row i is the 1-to-K coding of t_i. That is, if $\mathbf{x}_i \in C_k$ then $t_{ik} = 1$ and $t_{ir} = 0$ for $r \neq k$.

ML and softmax regression

The log-likelihood is then defined as

$$l = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik} \log y_{ik}$$

And the gradient is defined as

$$\nabla_{\overline{\mathbf{W}}} l = (\nabla_{\overline{\mathbf{w}}_1} l, \ldots, \nabla_{\overline{\mathbf{w}}_K} l)$$

where

$$\nabla_{\overline{\mathbf{w}}_k} l = \sum_{i=1}^{n} (t_{ik} - y_{ik}) \overline{\mathbf{x}}_i$$

Observe that the gradient has the same structure than in the case of linear regression and logistic regression