Machine learning

Probabilistic classification - discriminative models

Corso di Laurea Magistrale in Informatica

Università di Roma Tor Vergata

Prof. Giorgio Gambosi

a.a. 2023-2024

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

$$y(\mathbf{x}) = f(\mathbf{w}^T\mathbf{x} + w_0)$$

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

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

$$f(\mathbf{w}^T\mathbf{x} + w_0) = c$$

and

$$\mathbf{w}^T\mathbf{x} + w_0 = f^{-1}(y) = c'$$

(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 y 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 of y given $\mathbf{x}$, $p(y|\mathbf{x})$ belongs to the exponential family

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

2. for any $\mathbf{x}$, we wish to predict the expected value of $\mathbf{u}(y)$ given $\mathbf{x}$, that is $E[\mathbf{u}(y)|\mathbf{x}]$
3. $\boldsymbol{\theta}(\mathbf{x})$ (the natural parameter) is a linear combination of the features, $\boldsymbol{\theta}(\mathbf{x}) = \mathbf{w}^T\overline{\mathbf{x}}$

GLM and normal distribution

1. $y \in \mathbb{R}$, and $p(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu(\mathbf{x}))^2}{2\sigma^2}}$ 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 $\mathbf{u}(y) = y$
2. we wish to predict the value of $E[\mathbf{u}(y)|\mathbf{x}]$ as $y(\mathbf{x}) = E[y|\mathbf{x}]$, then

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

3. we assume there exists $\mathbf{w}$ such that $\theta_1(\mathbf{x}) = \mathbf{w}_1^T\overline{\mathbf{x}}$

Then, a linear regression results

$$y(\mathbf{x}) = \mathbf{w}_1^T\overline{\mathbf{x}}$$

GLM and Bernoulli distribution

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

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

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

$$p(y = 1|\mathbf{x}) = \pi(\mathbf{x}) = \frac{1}{1 + e^{-\theta(\mathbf{x})}}$$

3. we assume there exists $\mathbf{w}$ such that $\theta(\mathbf{x}) = \mathbf{w}^T\overline{\mathbf{x}}$

Then, a logistic regression derives

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

GLM AND CATEGORICAL DISTRIBUTION

1. $y \in \{1, \ldots, K\}$, and $p(y|\mathbf{x}) = \prod_1^K \pi_i(\mathbf{x})^{y_i}$ (where $y_i = 1$ if $y = i$ and $y = 0$ otherwise) is a categorical distribution with probabilities $\pi_1(\mathbf{x}), \ldots, \pi_K(\mathbf{x}))$ (where $\sum_{i=1}^K \pi_i(\mathbf{x}) = 1$): 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 $\mathbf{u}(y) = (y_1, \ldots, y_K)^T$ is the 1-to-K representation of y

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

$$p(y = i|\mathbf{x}) = E[u_i(y)|\mathbf{x}] = \pi_i(\mathbf{x}) = \pi_K(\mathbf{x}) e^{\theta_i(\mathbf{x})}$$

   Since $1 = \sum_{i=1}^K \pi_i(\mathbf{x}) = \pi_K(\mathbf{x}) \sum_{i=1}^K e^{\theta_i(\mathbf{x})}$, 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 there exist $\mathbf{w}_1, \ldots, \mathbf{w}_K$ such that $\theta_i(\mathbf{x}) = \mathbf{w}_i^T \overline{\mathbf{x}}$

GLM and categorical distribution

Then, a softmax regression results, with

$$
\begin{aligned}
y_i(\mathbf{x}) &= \frac{e^{\mathbf{w}_i^T\overline{\mathbf{x}}}}{\sum_{j=1}^K e^{\mathbf{w}_j^T\overline{\mathbf{x}}}} \qquad\qquad \text{if } i \neq K \\
y_K(\mathbf{x}) &= \frac{1}{\sum_{j=1}^K e^{\mathbf{w}_j^T\overline{\mathbf{x}}}}
\end{aligned}
$$

GLM AND ADDITIONAL REGRESSIONS

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

1. Assume $y \in \{0, \ldots, \}$ is a non negative integer (for example we are interested to count data), and $p(y|\mathbf{x}) = \frac{\lambda(\mathbf{x})^y}{y!}e^{-\lambda(\mathbf{x})}$ 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 $\mathbf{u}(y) = y$
2. we wish to predict the value of $E[\mathbf{u}(y)|\mathbf{x}]$ as $y(\mathbf{x}) = E[y|\mathbf{x}]$, then
$$y(\mathbf{x}) = \lambda(\mathbf{x}) = e^{\theta(\mathbf{x})}$$
3. we assume there exists $\mathbf{w}$ such that $\theta(\mathbf{x}) = \mathbf{w}^T\overline{\mathbf{x}}$

Then, a Poisson regression derives
$$y(\mathbf{x}) = e^{\mathbf{w}^T\overline{\mathbf{x}}}$$

GLM AND ADDITIONAL REGRESSIONS

1. Assume $y \in [0, \infty)$ is a non negative real (for example we are interested to time intervals), and $p(y|\mathbf{x}) = \lambda(\mathbf{x})e^{-\lambda(\mathbf{x})y}$ is an exponential distribution with parameter $\lambda(\mathbf{x})$: then, the natural parameter $\theta(\mathbf{x})$ is
$$\theta(\mathbf{x}) = -\lambda(\mathbf{x})$$
and $\mathbf{u}(y) = y$
2. we wish to predict the value of $E[\mathbf{u}(y)|\mathbf{x}]$ as $y(\mathbf{x}) = E[y|\mathbf{x}]$, then
$$y(\mathbf{x}) = \frac{1}{\lambda(\mathbf{x})} = -\frac{1}{\theta(\mathbf{x})}$$
3. we assume there exists $\mathbf{w}$ such that $\theta(\mathbf{x}) = \mathbf{w}^T\overline{\mathbf{x}}$

Then, an exponential regression derives

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

Discriminative approach

We could directly assume that $p(C_k|\mathbf{x})$ 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 y, which results into

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

where 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

- In the case of d features, logistic regression requires $d+1$ coefficients $w_0, \dots, w_d$ 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

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

  - one prior cla 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

Let us 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 $p_i = p(C_1|\mathbf{x}_i) = \sigma(\mathbf{w}^T\mathbf{x}_i)$.
Then, the likelihood of the training set targets $\mathbf{t}$ given $\mathbf{X}$ is

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

and the log-likelihood is

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

Maximum likelihood estimation

- It results

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

Maximum likelihood estimation

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

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

Maximum likelihood estimation

As a possible alternative, at each iteration only one coefficient in $\mathbf{w}$ is updated

$$
\begin{aligned}
w_k^{(j+1)} &= w_k^{(j)} + \alpha \frac{\partial l(\mathbf{w}|\mathbf{X},\mathbf{t})}{\partial w_k}\Big|_{\mathbf{w}^{(j)}} \\
&= w_k^{(j+1)} + \alpha \sum_{i=1}^{n} (t_i - \sigma((\mathbf{w}^{(j)})^T \overline{\mathbf{x}}_i)) x_{ik} \\
&= w_k^{(j+1)} + \alpha \sum_{i=1}^{n} (t_i - y(\mathbf{x}_i)) x_{ik}
\end{aligned}
$$

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 $\mathbf{\Sigma}$ 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 and GDA

- 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 $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_K)$ 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

$$p(\mathbf{T}|\mathbf{X}, \mathbf{W}) = \prod_{i=1}^{n} \prod_{k=1}^{K} p(C_k|\mathbf{x}_i)^{t_{ik}} = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\frac{e^{\mathbf{w}_k^T \overline{\mathbf{x}}_i}}{\sum_{r=1}^{K} e^{\mathbf{w}_r^T \overline{\mathbf{x}}_i}} \right)^{t_{ik}}$$

  where $\mathbf{X}$ is the usual matrix of features 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

$$l(\mathbf{W}) = \sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}\log\left(\frac{e^{\mathbf{w}_k^T\overline{\mathbf{x}}_i}}{\sum_{r=1}^{K} e^{\mathbf{w}_r^T\overline{\mathbf{x}}_i}}\right)$$

And the gradient is defined as

$$\frac{\partial l(\mathbf{W})}{\partial \mathbf{W}} = \left(\frac{\partial l(\mathbf{W})}{\partial \mathbf{w}_1}, \ldots, \frac{\partial l(\mathbf{W})}{\partial \mathbf{w}_K}\right)$$

ML and softmax regression

- It is possible to show that

$$\frac{\partial l(\mathbf{W})}{\partial \mathbf{w}_j} = \sum_{i=1}^{n} (t_{ij} - y_{ij})\overline{\mathbf{x}}_i$$

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

ML AND SOFTMAX REGRESSION: GRADIENT ASCENT METHOD

- Applying a gradient method to maximize the log-likelihood $l(\mathbf{w})$ requires using the gradient $\dfrac{\partial l}{\partial \mathbf{w}}$ to explore the dK-dimensional space of model coefficient values
- As an alternative, on-line gradient descent: at each iteration the ascent is performed only wrt to a cyclically selected coefficient $\mathbf{w}_k$, evaluating only the gradient $\dfrac{\partial l}{\partial \mathbf{w}_k}$ in a space of dimension d