Metody systemowe i decyzyjne

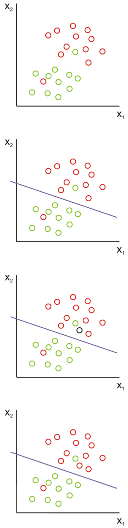
Zadanie regresji logistycznej

Szymon Zaręba

5 kwietnia 2017

Problem klasyfikacji

- **Zmienne wejściowe**, **atrybuty** (ang. *input variables*, *attributes*): $\mathbf{x} \in \mathcal{X}$
- **Zmienna wyjściowa**, **klasa**, **etykieta** (ang. *target variable*, *class label*): $y \in \{0, 1\}$ lub $y \in \{-1, 1\}$.
- **Problem:** dla $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ przewidzieć wartość klasy (etykietę) y dla nowego obiektu $\mathbf{x}$.
- Zgodnie z teorią decyzji wystarczy znać rozkład warunkowy $p(y|\mathbf{x})$, zatem chcemy go **modelować**.



Regresja logistyczna

- Model **regresji logistycznej** (ang. *logistic regression*):

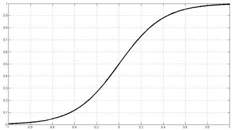
$$p(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}))$$

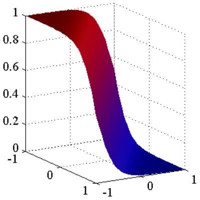
- **Funkcja sigmoidalna** (ang. *sigmoid function*):

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

- Predykcja

$$y^* = \begin{cases} 1, & \text{jeśli } p(y = 1|\mathbf{x}, \mathbf{w}) \geqslant \theta, \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$$





Funkcja wiarygodności

- Dane: $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, $\mathbf{y} = \{y_1, \ldots, y_N\}$.
- **Warunkowa funkcja wiarygodności** ($\sigma_n \equiv \sigma(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n))$):

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} \sigma_n^{y_n}(1-\sigma_n)^{1-y_n}$$

- Logarytm funkcji wiarygodności z minusem:

$$-\ln p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = -\sum_{n=1}^{N} \Big(y_n \ln \sigma_n + (1-y_n)\ln(1-\sigma_n) \Big)$$

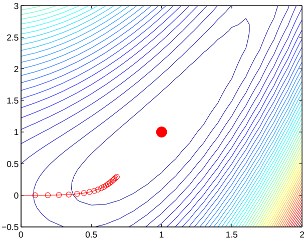
- Postać σ_n zależnej od parametrów $\mathbf{w}$ **nie pozwala** na analityczne rozwiązanie poprzez przyrównanie gradientu do zera.

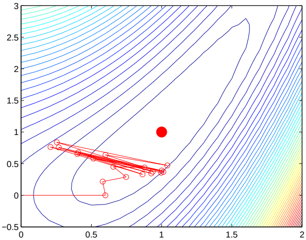
Algorytm gradientu prostego

- Pseudokod:

  Initialize **w**
  repeat
  $\quad \mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla_{\mathbf{w}} E(\mathbf{w})$
  until convergence

- Algorytm wrażliwy na dobór **parametru uczenia** α oraz **optima lokalne**.





Algorytm gradientu prostego

- Pseudokod:

  Initialize $\mathbf{w}$
  repeat
  $\quad \mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla_{\mathbf{w}} E(\mathbf{w})$
  until convergence

- Optymalizowana **funkcja celu** (podzielona przez N):

$$E(\mathbf{w}) = -\frac{1}{N} \sum_{n=1}^{N} \Big(y_n \ln \sigma_n + (1 - y_n) \ln(1 - \sigma_n) \Big)$$

- **Gradient** funkcji celu względem parametrów:

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = \cdots$$

Stochastyczny algorytm gradientu prostego

- Pseudokod:

  Initialize **w**
  repeat
      for $n = 1$ to N **do**
          $\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla_w E_n(\mathbf{w})$
      end for
  until convergence

- **Przybliżenie** funkcji w n-tym kroku:

$$E_n(\mathbf{w}) = -y_n \ln \sigma_n - (1 - y_n) \ln(1 - \sigma_n)$$

- **Gradient** powyższej funkcji:

$$\nabla_{\mathbf{w}} E_n(\mathbf{w}) = \cdots$$

Stochastyczny algorytm gradientu prostego

- SGD zbiega zazwyczaj **dużo szybciej** niż algorytm gradientu prostego.
- Często w pojedynczym kroku jest niestabilny i pojawiają się silne **drgania** (*ang. jitterings*).
- W celu redukcji tego efektu zamiast pojedynczej obserwacji, wykorzystuje się **małe paczki** danych (*ang. mini-batch*) o wielkości M.
- Wtedy dla pojedynczej paczki MB funkcja ma postać:

$$E_{MB}(\mathbf{w}) = -\frac{1}{M} \sum_{n \in MB} \big(y_n \ln \sigma_n + (1 - y_n) \ln(1 - \sigma_n) \big)$$

Regularyzacja

W rozważanym problemie wprowadzamy regularyzację ℓ_2 na parametry (oprócz wyrazu wolnego):

$$L_\lambda(\mathbf{w}) = L(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}_{-0}\|_2^2$$

gdzie $\mathbf{w}_{-0} = (w_1, \ldots, w_{D-1})^{\mathrm{T}}$ jest wektorem wag bez pierwszego parametru, tzw. wyrazu wolnego (ang. *bias*) w_0, $\lambda > 0$ oznacza współczynnik regularyzacji.

Selekcja modelu

W rozważanym problemie mamy do czynienia z dwoma wielkościami, których nie wyuczamy w oparciu o dane:

- wartość progowa klasyfikacji θ
- wartość współczynnika regularyzacji λ

W celu oceny modelu w oparciu o wybrane wielkości obu parametrów, stosować będziemy miarę **F-measure**:

$$F(\mathcal{D}_{val}; \theta, \lambda, \mathbf{w}) = \frac{2TP}{2TP + FP + FN}$$