ТЕМА 5
КЛАСИФІКАЦІЯ НА ОСНОВІ ТЕОРІЇ БАЙЕСА

Надія І. Недашківська n.nedashkivska@gmail.com

Постановка задачі для байесівської класифікації

$X = \{x_1, x_2, \ldots, x_m\}$ - незалежні змінні (атрибути)

$y \in Y$ - залежна змінна $X \times Y$ - ймовірнісний простір з щільністю розподілу $p(x, y)$

Дано: Множина об'єктів $T = \{t_1, t_2, \ldots, t_n\}$ навчальна вибірка (тестові приклади)

$t_i \rightarrow (X, y)$

Знайти:
функцію класифікації $f : X \rightarrow Y$ з мінімальною імовірністю помилки

$$p(x, y) = p(x)p(y \mid x) = p(y)p(x \mid y)$$

Максимум апостеріорної імовірності (maximum a posteriori, MAP)

$$f(x) = \operatorname*{argmax}_{y \in Y} p(y \mid x) = \operatorname*{argmax}_{y \in Y} \frac{p(x \mid y)p(y)}{p(x)} = \operatorname*{argmax}_{y \in Y} p(x \mid y)p(y)$$

Оптимальний байесівський класифікатор

Області, отримані в результаті класифікації:

$$X_s = \{x \in X \mid f(x) = s\}, \quad s \in Y$$

Помилка: об'єкт x класу y потрапляє в $X_s,\ s \neq y$.

Штраф за помилку відомий: $u_{ys} \geq 0,\ \forall (y,s) \in Y \times Y$

Середній ризик: математичне сподівання штрафу для f:

$$R(f) = \sum_{y \in Y} \sum_{s \in Y} u_{ys} p(X_s, y)$$

Якщо відомі $p(y),\ p(x \mid y)$, то мінімальний середній ризик має байесівський класифікатор*

$$f(x) = \arg\min_{s \in Y} \sum_{y \in Y} u_{ys} p(x \mid y) p(y)$$

* http://www.machinelearning.ru/

Оптимальний байєсівський класифікатор

Якщо відомі $p(y)$, $p(x \mid y)$, $u_{ys} = u_y$, $u_{yy} = 0$, то мінімальний середній ризик має класифікатор*

$$f(x) = \arg\max_{y \in Y} u_y p(x \mid y) p(y)$$

Підзадача 1: Дано: навчальна вибірка $T = \{t_i\}$, $t_i \to (X, Y)$.

Знайти: оцінки для $p(y)$, $p(x \mid y)$, $y \in Y$ за вибіркою.

Підзадача 2: Дано: апріорні імовірності $p(y)$, функції правдоподібності $p(x \mid y)$, $y \in Y$.

Знайти: класифікатор $f : X \to Y$, який мінімізує середній ризик.

* http://www.machinelearning.ru/

Оцінювання $p(y),\ p(x \mid y),\ y \in Y$ за вибіркою

Оцінювання апріорних імовірностей:

$$\hat{p}(y) = \frac{|T_y|}{|T|} \qquad T_y = \{x_i \in X :\ y_i = y\} \qquad y \in Y$$

Оцінювання функцій правдоподібності $p(x \mid y)$ **:**

- Параметричне оцінювання:

$$\hat{p}(x \mid y) = \varphi(x, y, \theta)$$

- Оцінювання (розділ) суміші розподілів:

$$\hat{p}(x \mid y) = \sum_{k=1}^{q} \omega_k \varphi(x, y, \theta_k) \qquad q << n$$

- Непараметричне оцінювання

Наївний байєсівський класифікатор *Naive Bayes*

Припущення: нехай $x_1, x_2, \ldots, x_m$ незалежні у сукупності.

Відомі щільності $p(x_1 | y),\ p(x_2 | y),\ \ldots,\ p(x_m | y)$, $y \in Y$.

Тоді m-вимірна щільність – це добуток 1-вимірних щільностей:

$$p(x | y) = p(x_1 | y) \cdot p(x_2 | y) \cdot \ldots \cdot p(x_m | y) \qquad y \in Y$$

Тоді класифікатор з мінімальним середнім ризиком:

$$\boxed{f(x) = \arg\max_{y \in Y} u_y p(x | y) p(y) = \arg\max_{y \in Y} u_y p(y) \prod_{i=1}^{m} p(x_i | y)}$$

$$f(x) = \arg\max_{y \in Y} \left(\ln(u_y \hat{p}(y)) + \sum_{i=1}^{m} \ln \hat{p}(x_i | y) \right)$$

нехай

$u_y = 1$

Алгоритм *Naive Bayes*: приклад

Чи виграє “Динамо” при наступних умовах (вектор *x*):

- x_1=в гостях,
- x_2= суперник нижче в турнірній таблиці,
- x_3=температура в нормі,
- x_4=дощу немає

Такого прикладу немає в навчальній вибірці.

$$f(x) = \arg\max_{y \in Y} p(y \mid x) = \arg\max_{y \in Y} p(y) \prod_{i=1}^{m} p(x_i \mid y)$$

Треба розрахувати дві апостеріорні імовірності класів:

p(перемога=так | *x*)

p(перемога=ні | *x*)

Приклад: результати попередніх ігор

Де грає	Суперник	Температура	Дощ	Перемога
Вдома	Вище	Висока	Так	Ні
Вдома	Нижче	Норма	Ні	Так
В гостях	Нижче	Норма	Так	Так
В гостях	Нижче	Висока	Так	Ні
Вдома	Вище	Висока	Ні	Так
Вдома	Нижче	Висока	Так	Ні
В гостях	Нижче	Висока	Ні	Ні
В гостях	Вище	Норма	Ні	Так

Алгоритм *Naive Bayes*: приклад

Шукані апостеріорні імовірності класів:

p(перемога=так | *x*) = *p*(перемога=так) *
p (x1=в гостях | перемога=так) *
p (x2=нижче | перемога=так) *
p (x3=норма | перемога=так) *
p (x4=ні | перемога=так) / p(*x*)

p(перемога=ні | *x*) = *p*(перемога=ні) *
p (x1=в гостях | перемога=ні) *
p (x2=нижче | перемога=ні) *
p (x3=норма | перемога=ні) *
p (x4=ні | перемога=ні) / p(*x*)

Алгоритм *Naive Bayes*: приклад

Функції правдоподібності $p(x_1 | y),\ p(x_2 | y),\ \ldots,\ p(x_m | y)$**,** $y \in Y$**:**

p(x$_1$=в гостях | перемога=так)=2/4

p(x$_1$=в гостях | перемога=ні)=2/4

p(x$_2$=нижче | перемога=так)=2/4

p(x$_2$=нижче | перемога=ні)=3/4

p(x$_3$=норма | перемога=так)=3/4

p(x$_3$=норма | перемога=ні)=0/4

p (x$_4$=ні | перемога=так) =3/4

p (x$_4$=ні | перемога=ні) =1/4

Апріорні імовірності класів $p(y)$**,** $y \in Y$ **:**

p(перемога=так)=4/8

p(перемога= ні)=4/8

Алгоритм *Naive Bayes*: приклад

Шукані апостеріорні імовірності класів:

p(перемога=так | x) = 2/4* 2/4* 3/4* 3/4* 4/8 / p(x)

p(перемога=ні | x) = 2/4* 3/4 * 0 * 1/4* 4/8 / p(x)

p(перемога=так | x) > p(перемога=ні | x)

Тому при заданих умовах матч буде виграно.