Распознавание образов

12.02.14

Классификация или отнесение объектов к одному из классов

Объекты, кот. классиф-ся наз. образами

Хар-ки объекта — это признаки

Необх. по признакам отнести объект к одному из классов на основе алгоритма — этот процесс наз. классификацией

Литература:

1. Местецкий "Классификация"
2. Воронцов → ШАД (школа анализа данных) → лекции (есть ссылка на конспект)
3. Николенко (ПОМИ) — кафедра мат. логики, уделяют внимание алгоритмической стороне вопроса — см. курс лекций.

Признаки бывают:

1. Двоичные (0 или 1) (да или нет)
2. Номинальные — конечное число неупоряд. значений.
3. Порядковые (ординарные) т.е. их можно упорядочить, но они не числовые хар-ра.
4. Числовые

Классиф-ия → с обучением — дано несколько объектов проклассиф-ых заранее на основе них м. принимать данн. реш-ия
→ без обучения — если нет классов, но необх. классиф-ть — задача кластерного анализа

Байесовские методы классификации

$(x_1, y_1) \ldots (x_n, y_n) \sim f(x, y)$ — см. бел.

где y принимает конечное число знач-ий.

x_i — признаки

В действ-ти, мы наблюдаем $x_1, \ldots, x_n$

Прогнозируем Y.

Бинарная классификация

$Y = 0$ или 1

Классификатор $g(x) = \begin{cases} 0 \\ \text{или} \\ 1 \end{cases}$ $\eta(x) = P(Y = 1 \mid X = x)$

$Y = \begin{cases} 0, & \text{с вер-тью } 1 - \eta(x) \\ 1, & \text{с вер-тью } \eta(x) \end{cases}$ $0 \le \eta(x) \le 1$

(при фикс. X, м. задавать Y)

Ошибка классиф-ии: ~~$P(g(x) \neq y \mid X)$~~ $P(g(X) \neq Y)$

Байесовский классификатор

$g^*(x) = \begin{cases} 1, & \eta(x) \ge \frac{1}{2} \\ 0, & \eta(x) < \frac{1}{2} \end{cases}$ $g = \eta$

$\boxed{T}$ $P(g^*(X) \neq Y) \le P(g(X) \neq Y)$

Док-во: $P(g(X) \neq Y) = E_x E(\mathbb{1}_{\{g(X) \neq Y\}} \mid X = x) \Rightarrow$

$\Rightarrow$ Дост-но миним-ть мат. ожидание индикатора $E(\mathbb{1}_{\{g(x) \neq y\}} \mid X = x) =$

$= 1 - E(\mathbb{1}_{\{g(x) = y\}} \mid X = x) =$

$= 1 - (P(g(x) = 1, Y = 1 \mid X = x) + P(g(x) = 0, Y = 0 \mid X = x)) =$

$= 1 - (\mathbb{1}_{\{g(x) = 0\}} \underbrace{P(Y = 0 \mid X = x)}_{1 - \eta(x)} + \mathbb{1}_{\{g(x) = 1\}} P(Y = 1 \mid X = x)) =$

$= 1 - (\mathbb{1}_{\{g(x) = 0\}} (1 - \eta(x)) + \mathbb{1}_{\{g(x) = 1\}} \eta(x)) =$

$= P(g^*(x) \neq Y \mid X = x) - P(g(x) \neq Y \mid X = x) =$

$= (\mathbb{1}_{\{g(x) = 0\}} - \mathbb{1}_{\{g^*(x) = 0\}})(1 - \eta(x)) +$

$+ (\mathbb{1}_{\{g(x) = 1\}} - \mathbb{1}_{\{g^*(x) = 1\}}) \eta(x)$

Заметим, $\mathbb{1}_{\{g(x) = 1\}} = 1 - \mathbb{1}_{\{g(x) = 0\}}$

$(\mathbb{1}_{\{g^*(x) = 1\}} - \mathbb{1}_{\{g(x) = 1\}})(1 - \eta(x)) + (\mathbb{1}_{\{g(x) = 1\}} -$
$- \mathbb{1}_{\{g^*(x) = 1\}}) \eta(x) =$

$= (\mathbb{1}_{\{g^*(x) = 1\}} - \mathbb{1}_{\{g(x) = 1\}})(1 - 2\eta(x)) \le 0$

$1 - 2\eta(x) > 0 \Rightarrow \mathbb{1}_{\{g^*(x) = 1\}} = 0$ $\square$

2

Общий случай:

$(X_1, Y_1), \dots, (X_K, Y_K)$ $\quad Y_i = \begin{cases} 1 \\ \vdots \\ K \end{cases}$

$P(Y_i = l \mid X_i = x) = \eta_i(x)$ $\quad \sum_i \eta_i(x) = 1$

min $P(g(X) \neq Y)$

Оптимальный классиф-ор: $\hat{l} = \arg\max_l \eta_l(x)$

Разделение смеси вероятностных распределений:

$p(x) = \sum_{l=0}^{K} \eta_l \cdot p_l(x)$

Ответ: $P(Y = l \mid X = x) = \dfrac{\eta_l p_l(x)}{\sum \eta_i p_i(x)}$

$\hat{p} = \arg\max_l \eta_l p_l(x)$

Классификация для смеси нормальных распределений. Дискриминантный анализ.

$p(x) = \sum \eta_i N(x \mid \mu_i, \Sigma_i)$ вер-то появления того или иного класса; корреп. матр.; параметры норм. распр-ия

Необх. найти:

$$\arg\max \eta_i N_i(x \mid \mu_i, \Sigma_i)$$

// Если есть обучающая выборка, то оцениваем μ_i и Σ_i, если нет, то есть ем-алгоритм //

$\hat{l} = \arg\max \ln \eta_l + \ln N_l(x, \mu_l, \Sigma_l)$

(*) $\ln \eta_l - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| -$

$\ln \eta_l - \frac{1}{2} \ln$ X - то (.), в кот. мы классифицируем

$\ln(\exp \dots)$

$-\frac{1}{2}(x - \bar{x}_l) S_l^{-1} (x - \bar{x}_l) \to \arg\max$

оценка μ_l; оценка Σ_l

расстояние Махаланобиса

Дискриминант Фишера

$\exists \quad \Sigma_1 = \Sigma_2 = \dots = \Sigma_n$

Тогда м. построить общую оценку для дисперсии S.

$$(*): \underbrace{\ln \eta_\ell - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma|}_{=:\, c_\ell} - \frac{1}{2}(x-\bar{x}_\ell)' S_\ell^{-1}(x-\bar{x}_\ell)$$

Тогда $x S^{-1} x$ сокращается т.к. не зависит от ℓ

$$\arg\max \underbrace{c_\ell + \overbrace{\bar{x}_\ell S^{-1} \bar{x}_\ell}^{x' S^{-1} \bar{x}_\ell} - \frac{1}{2}\bar{x}_\ell S^{-1} \bar{x}_\ell}_{x S^{-1} \bar{x}_\ell} \leftarrow \text{лин. фция от } x.$$

Дискриминант Фишера

// Находим гиперпл-ть на кот достигается max и там осущ-ся классификация //

0.02.14. $p(x) = \sum_{i=1}^{K} \omega_i p_i(x, \theta_i)$ — модель, где K — число классов

распред-ие внутри класса.

Предположим, что обучающей выборки нет. => придумаем EM-алгоритм (expectation, maximization) — состоит из этих 2^K шагов

Задача моделирования смеси распределения:

1. (Шаг) λ_1: $P(\lambda_i = i) = \omega_i$

2. Моделируем x_i с плотн-ю распр-ия $p_i(x)$

4 // Разыграли номер, после самму сл. в.

$(\lambda_1, x_1), \dots, (\lambda_n, x_n)$

$\hat{w}_i = \frac{1}{n}\sum_{s=1}^{n} \mathbb{1}\{\partial_s = i\}$ – оценка w_i. (*)

∃ $X_1 \dots X_n$ – сл. величины, а $\partial_1, \dots, \partial_n$ не наблюдаются
Хотим проследить сл. в. $\partial_1, \dots, \partial_n$. как построить $\hat{w}_i$ как оценку

$P(\partial = i \mid X = x) =$ //по т. Байеса// $= \frac{w_i p_i(x)}{\sum_{\tau=1}^{k} w_\tau p_\tau(x)} = g_i(x)$

После того как ~~оценили~~ по модели оценим ∂_i, м оценить $\hat{w}_i$ (*)

∀ рандомизация ↑ дисперсию оценку => индикаторы в (*) заменяем их мат. ожид-ми в (*)

$E(\hat{w}_i \mid X_1, \dots, X_n) = \frac{1}{n}\sum_{s=1}^{n} g_{si}(X_s) =: \hat{w}_i$ – обозн. новая оценка

// $\approx \frac{1}{n}\sum_{s=1}^{n} \underbrace{E(\mathbb{1}\{\partial_s = i\} \mid X_s)}_{g_i(X_s)}$ // то научились просто оценивать $\hat{w}_i$ → м написать алг. мод-ия.

Алгоритм:

1° Берем нач. приближение w_{i_0}, p_{i_0}, $l=0$

2° E-шаг Оцениваем $\hat{w}_{il} = \frac{1}{n}\sum_{s=1}^{n} \frac{\hat{w}_{il-1} p_i(X_s, \hat{\theta}_{il-1})}{\sum \hat{w}_{il-1} p_i(X_s, \hat{\theta}_{il-1})}$

делаем оценку некот. мат. ожидания

3° M-шаг (находим оценку $\hat{\theta}_{il}$ max правдоподобия для θ_i при условии, что $w_i = \hat{w}_{il}$) – известны.

★

P.S. Оц. max правд. напишем не совсем в очевидной форме, обоснуем ее позже

(*) $\sum g_{il}(X_s) \frac{p_{i\theta}(X_s, \theta_i)}{p_i(X_s, \theta_i)} = 0$ "+" Решаем отдельно $\forall i$, т.е. мож-но решать сист. урий

$g_{il}(x) = \frac{\hat{w}_{il} p_i(x, \theta_{il-1})}{\sum \hat{w}_{jl} p_i(x, \theta_{j,l-1})}$ ← при фикс. i – зам. св-во метода

Именно это нам нужно сейчас обосн.

Общее обоснование (*):

$$\max \sum_{s=1}^{n} \ln\left(\sum_{j=1}^{k} \hat{w}_j p_j(x, \theta_j)\right) \to \text{продифф. по } \theta_i$$

Получим: $\sum_{s=1}^{n} \left(\hat{w}_{j,l-1} / \sum w_{j,l-1} p_j(x_s, \theta_j)\right) \cdot p_{i\theta}(x_s, \theta_i) = 0$

// Потом выведем это для $N(a, \sigma^2)$ //

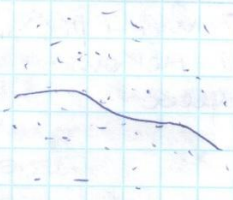
Умножим на $p_i(x_s, \theta_i) / p_i(x_s, \theta_i)$. Тогда:

$$\sum \underbrace{\frac{\hat{w}_{i,l-1} \cdot p_i(x_s, \theta_i)}{\sum w_{j,l-1} p_j(x_s, \theta_j)}}_{=: g_{i,l-1}(x_s)} \cdot \frac{p_{i\theta}(x_s, \theta_i)}{p_i(x_s, \theta_i)}$$

$$0 = \sum_{s=1}^{n} g_{i,l-1}(x_s) \cdot \frac{p_{i\theta}(x_s, \hat{\theta}_{i,l})}{p_i(x_s, \hat{\theta}_{i,l})} = 0$$

Замечание: м. показать, что оценки E-шага явл. оц. max правд-ия w_i при усл-ии что θ_i – фиксир. Таким образом мы находим ОМП поочередно параметры θ и w фиксируя

Чем хороший алгоритм? один из них



Можно взять для каждого класса смесь нормальных распределений. – это повлияет на границу

Вставка (см. посл. Каштанова)

Смесь нормальных распр-ий

$$p(x) = \sum w_i N(\mu_i, R_i)$$

↑ плотность норм. распр.

Шаг 1: нач. приближ-ие w_{i0}, μ_0, R_{i0}.

Шаг 2: $g_{il}(x) = \dfrac{\hat{w}_{il} N(\hat{\mu}_{i,l-1}, \hat{R}_{i,l-1})}{\sum \hat{w}_{j,l-1} N(\hat{\mu}_{j,l-1}, \hat{R}_{j,l-1})}$

$w_{il} = \frac{1}{n} \sum_s g_{i,l-1}(X_s)$

Шаг 3: $\hat{\mu}_{ie} = \sum g_{ie}(x_s) \cdot \vec{x}_s$ // μ_e и x_s – вектора

$$\hat{R}_{ie} = \sum_{s=1}^{n} g_{ie}(x_s)(\vec{x}_s - \hat{\mu}_{ie})(\vec{x}_s - \hat{\mu}_{ie})^T$$

$\hat{w}_{ie} = \frac{1}{n}\sum^{n} g_{ie}(x_s)$ – верно – будет в общем случае и в сл. норм. распр-я

$w_{ie} = k/N$

Непарам. ~~оценивание в задачах оценивания нормального распределения~~ плотности классификации

Применение методов непараметрического оценивания в задачах классификации (Классификация с обучением)

$$p(x) = \sum_{j=1}^{n} w_j p_j(x)$$

Если есть обучающая выборка: $(x_1, y_1), \dots, (x_n, y_n)$, то мы взять относительные частоты $\hat{w}_j = \frac{1}{n}\sum$

$$\hat{w}_j = \frac{1}{n}\sum_{s=1}^{n} \mathbb{1}\{y_s = j\}$$ как оценки w_j; после выделить подвыборки $x_{s_1}, \dots, x_{s_\ell}$ т.ч. $y_{s_\ell} = j$ и для них непараметрически оценить плотность $p_j(x)$ и задача будет решена

P.S. осн. проблема – вектор признаков X многомерный, а мы его в добавок разбиваем на кусочки.

Выход:

Ядерная оценка: $\hat{f}_n(x) = \frac{1}{nh}\sum_{s=1}^{n} K\left(\frac{\rho(x_s, x)}{h}\right)$,

где $\rho(x_s, x)$ – расстояние м/у x_s и x.

Возникает вопрос выбора h.

Идея: выбрать h т.ч. что $\#\{x_s : \rho(x_s, x) < h\} = k$ – фикс.

Если $K(x) = \mathbb{1}\{|x| \le 1\}$, то мы получаем в кач-ве оценки:

$$\hat{f}_n(x) = \frac{k}{n\rho(x, x^{(k)})},$$ где $x^{(k)}$ – это k-ое удаленное наблюдение от x.

$$\rho(x, x^{(1)}) \le \rho(x, x^{(2)}) \le \dots \le \rho(x, x^{(k)})$$

На основе такой оценки построен алгоритм k-ближ. соседей

Алгоритм k-ближайших соседей.

1 имеется наблюдение: $(X_1, Y_1), \dots, (X_n, Y_n)$
Хотим найти классификацию для X_{new}.
Берем для него k ближайших $X^{(1)}, \dots, X^{(k)}$
$\rho(X_{new}, X^{(1)}) \le \dots \le \rho(X_{new}, X^{(k)})$

$y^{(i)}$ – категория $X^{(i)}$
$\partial_i = \sum_{s=1}^{k} \mathbb{1}\{y^{(s)} = i\}$; Категория $X_s = \arg\max \partial_i$
// т.е. берем то, что чаще всего встречается //

Критерий ог сильно зависит от выбора расстояний и от k. (чаще $k=1$)

27.02.14. Параметрические методы классифик.

$(X_1, Y_1), \dots, (X_n, Y_n)$ – обучающая выборка
$\{X_i\}$ – признаки, $\{Y_i\}$ – образы (объекты)
$Y_i = \{1, -1\}$

Опр. Ф-ия дискриминации $f(x, w)$ параметры $w \in \mathbb{R}^d$ (*
Опр. Дискриминатор $a(x) = \operatorname{sign} f(x, \hat{w})$
оценка

Мы построить эмпирич. риск: $R(X^{(n)})$
+ введем фию $a(x, w) = \operatorname{sign} f(x, w)$. Тогда:
$X^{(n)} = (X^{(1)}, \dots, X^{(n)})$: $R(X^{(n)}) = \sum_{i=1}^{n} \mathbb{1}\{f(x_i, w) y_i < 0\}$ (*)
эмпирический риск
Нам необх. найти W т.ч. (*) → min.
Проблемы: 1) фии, кот. мин-ся неадекватно
2) м. ∄ ог. богатое семейство параметров
тогда хорошо min эмпир риск, но фия
// – проблема переобучения (переподгонки) //
$f(x, w)$ м. иметь нерациональный вид

8

Поэтому в реальных задачах обучающая выборка разиб-ся на 2 части: по первой строим $f(\hat{w})$, по 2ой проверяем, не произошёл ли эффект переобучения.

Т.о. в задачах классификации не только строим классификатор, но и проверяем его.

Опр: Величина $f(x_i, w) y_i$ наз. отступом.
$M := f(x_i, w) y_i$

Как бороться с $\mathbb{1}$ в формуле (*)?
(?): $\mathbb{1}$ заменяем на $\mathcal{L}\{f(x_i, w) y_i\}$ = непрерывная фция. $\mathcal{L}(M)$

Пр: $\mathcal{L}(M) = \frac{1}{1+e^{-M}}$ – логистическая регрессия

$(1-M)_+ = \max(1-M, 0)$
↑ SVM – support vector machines

e^{-M} – ADA boosting

(*) заменяем на

(**) $Q = \sum \mathcal{L}(M_i) = \sum_{i=1}^{n} \mathcal{L}(y_i f(x_i, w))$ – это отступ M – min
// хотим получить ОМП //

$\exists\ \mathcal{L}(y_i f(x_i, w)) \approx -\ln p(x_i, y_i, w)$, где
p – плотность распр-ия x_i, y_i

Тогда задача (**) эквив-на задаче нахождения ОМП для $p(x, y, w)$

Т.е. $Q = -\sum_{i=1}^{n} \ln p(x_i, y_i, w)$

Дискрим. анализ для смеси распр-ия и логист. регрессии допускают также вероятностную интерпретацию

Классификация на основе логистической регрессии

Есть наблюд-ия $(x_1, y_1), \dots, (x_n, y_n)$ $\quad y_i = \{1, -1\}$

М написать совместн плотность распреде:

$p(x, y) = p(y|x) \cdot p(x)$, $\quad y = \{1, -1\}$

Предположим, что $p(1|x) = \exp\{c(\delta)\langle w_1, x\rangle + h(\delta, x) + r(\delta, w_1)\}$

$p(-1|x) = \exp\{c(\delta)\langle w_{-1}, x\rangle + h(\delta, x) + r(\delta, w_{-1})\}$

Опт δ имеет след. вид $\prod p(y_i, x_i, w_{y_i})$

$\sum \ln p(y_i, x_i, w_{y_i}) = \sum \ln p_w(y_i|x_i) + \underbrace{\sum \ln p(x_i)}_{\text{не зависит от } w} \to \max$

$\Rightarrow$ В результ. остается найти $\underbrace{\sum \ln p_w(y_i|x_i)}_{(\Leftrightarrow)}$ — заменим на $\dfrac{p(y_i|x_i)}{p(1, x_i) + p(-1, x_i)}$

// знаменатель не зависит от параметра и может быть взят из $p(x)$

$$(\Leftrightarrow) \quad \sum \ln \frac{p(y_i|x_i)}{p(1|x_i) + p(-1|x_i)} \to \max$$

$$y = 1: \quad \frac{\exp\{c(\delta)\langle w_1, x\rangle + r(\delta, w_1)\}}{\exp\{c(\delta)\langle w_1, x\rangle + r(\delta, w_1)\} + \exp\{c(\delta)\langle w_{-1}, x\rangle + r(\delta, w_{-1})\}} =$$

$$= \frac{1}{1 + \exp\{-y\, c(\delta)\langle w_1 - w_{-1}, x\rangle + r(\delta, w_{-1}) - r(\delta, w_1)\}} =$$

Добавим к x еще одну коорд $(x, 1)$

$$= \frac{1}{1 + \exp\{-y\, c(\delta)\langle w_1 - w_{-1}, x\rangle\}} \quad (*)$$

В случае $y = -1$ $\quad \dfrac{p(-1|x)}{p(1|x) + p(-1|x)}$ — будет $= (*)$

$\sigma(t) = \frac{1}{1+e^{-t}}$ – сигмоидальная функ

$Q = \sum \ln \sigma(y_i \langle w, x_i \rangle) \to \max$ $w = w_1 - w_{-1}$

/ необх. найти экстремум //

$\sigma(t)$ присутствует во всех нейронных сетях.
Она замечательна тем, что
$\sigma'(t) = \sigma(t)(1 - \sigma(t)) = \sigma(t)\sigma(-t)$

Градиентный метод:

$w_k = w_{k-1} - \eta \nabla Q$ // метод оценки w //

$\nabla Q = \sum_{i=1}^{n} \sigma(-y_i \langle w, x_i \rangle) y_i x_i$.

Метод стохастического градиента

1. Берем нач. приближение $w_1, \ldots, w_d$
2. Вычисляем Q
3. Рыбырываем случайно № $i =: \tau$
   $P(\tau = \tau) = 1/n$, $1 \le \tau \le n$ (т.е. равномерно)
4. Пересчитываем $w := w - \eta \sigma(-y_\tau \langle w, x_\tau \rangle) y_\tau x_\tau$
5. Вычисляем $\varepsilon = \ln \sigma(-y_\tau \langle w, x_\tau \rangle)$
6. Берем $Q = \frac{n-1}{n} Q + \varepsilon$ // Посчитать м.о. //
7. Останавливаемся, если знач. w и Q стабилизировались, иначе до то 3°.

Проблемы, связанные со сходимостью метода:
(Когда возникают?)

1. Число парам-ов (w) м.б. >, т.е. не сопоставимо с объемом обуч-щей выборки

2. $\{x_i\}$ м.б. сильно коррелированы м/у собой

3. Некот. признаки м. нести мало инф-ции
   => коэф-ты w б. оцениваться плохо