

Machine Learning

- ① Strong vs Weak: one area / human-like
 ANI AGI

Rules based vs ML

Knowledge - patterns. Gained. Help to predict and solve.

- ② ML - computers work w/o expl. programmed
 Supervised: regression, classif., forecasting
 Un - II -: clusterisation, dim. reduction
 Semi-sup.: some data is unlabeled.

X - объекты, Y - gen. others. $g: X \rightarrow Y$ - описание
 функции. Хорошее narrow g , too. шир. g .
 (x, y) - примеры $(y = g^*(x))$ на всем X
 f - признак: $f: X \rightarrow Df$

Од схема - $f(f_i)$ - описание, параметрическое, нелинейное
 нет - разнородные,
 $(f_1(x) \dots f_n(x))$ - гибкое описание X
 $\|f_i(x_j)\|$ - матрица объектов-признаков F

$Y = \{1 \dots M\}$ - классы. $Y = \{0, 1\}^M$ - квад. вектор
 $Y = \mathbb{R}^n$ - реальное M непрер. классов

$A = \{g(x, \theta) | x \in X, \theta \in \Theta\}$, Θ - диапазон параметров.

$\underbrace{\text{модель}}$ $g(x, \theta) = \sum_{i=1}^n \theta_i f_i(x)$ - лин. модель

Мног. обучение - $M: (\overbrace{X \times \mathbb{R}}^{\text{sign}})^E \rightarrow A$ - θ -апп.

$X^E = (x_i, y_i)_{i=1}^E$ - библиотека
 Loss function (функция потерь) \mathcal{L}

$\mathcal{L}(g, x) = 0 \iff$ корректно
 Гауссова концепция $(Q(g, X^E)) = \sum_{i=1}^E \mathcal{L}(g, x_i)$
 (математическое ожидание)

$\mathcal{L}(g, x) = [g(x) - y^*(x)]^2$ - квадратич.
 $|g(x) - y^*(x)|$ - отклонение (важн. для лин.)

ERM - empiric risk minimization: $\mu(X^t) = \text{argmin}_{f \in \mathcal{F}} Q(f, X^t)$
 виды задач: классификация, регрессия, ранжирование, кластеризация.

$$(3) \quad a(y, X^t) = \underset{y \in Y}{\text{argmax}} \Gamma_y(\mu, X^t); \quad \Gamma_y(y, X^t) = \sum_{i=1}^n [y_i^i = y] w_i; y$$

многоклассовая классификация
линейное обучение
без взвешивания по
вероятности

- NN: обучение - зан. алгоритм
 - не каскадизируемое
 - чувствитель к непривидимым / выбросам

$$\bullet kNN: w(i, n) = [\underset{\text{правильна}}{i \leq k}] \underset{\text{через}}{\text{Leave-one-out}} \underset{\text{уменьшает}}{\text{участие}} \underset{\text{непр. классиф!}}{\text{некорректно}}$$

$$LOO(k, X^t) = \sum_{i=1}^t [\underset{k \neq i}{a(x_i, X^t \setminus x_i)}] \rightarrow \min$$

$$\bullet kNN \text{ weighted: } w(i, n) = [\underset{\text{равн. проба}}{i \leq k}] w_i, \quad a(y, X^t, k) = \underset{\text{где}}{\text{argmax}} \sum_{i=1}^k [y_i^i = y] w_i$$

- Все есть хранилище X^t (можно хранить, не смотря на ...)

- $O(t)$ много. Можно $O(\ln t)$, но есть ...

- Параллельное обуч. К-дерево - листы из $[0, \infty)$

$$w(i, n) = K \left(\frac{1}{h} g(y_i, x_n^{(i)}) \right)$$

h можно брать перенесением: К-членами
(логич. нейрон. грави)
на $[0, 1]$

h - нач. число, при котором K сажает нач. числа. листа.
 $h(a) = g(a, x_n^{(k+1)})$.

- Hold-out: $T^e = T^t \cup T^{t-e}$

$$HO = Q(\mu / T^t), T^{t-e}) \xrightarrow{\text{train test}} \min$$

- Complete $L-V$ - split all possible
- $CCV_t = \frac{1}{C_t^{L-t}} \sum_{T^t} Q(\mu / T^t), T^{L-t}) \rightarrow \min$
- k-fold CV
- $CV_k = \frac{1}{k} \sum_{i=1}^k Q(\mu / T^i / F_i), F_i) \rightarrow \min$
- $t \times k$ -fold - split T^t t times randomly
- LOO (leave-one-out)
 - train - T^{t-1} , test = $\{x_i\}$.
 - $LOO_t = \frac{1}{t} \sum_{i=1}^t Q(\mu / T^{t-1} / p_i), p_i) \rightarrow \min$
- Random, отбор

Интуиция - близок к оценкам ~ 6 раз + те же классы
 Переизучение / пересечение - можно удалять
 Выброс - избыточные - в этом случае

$$M(x_i) = \Gamma_{y_i}(x_i) - \max_{y \in Y \setminus \{y_i\}} \Gamma_y(x_i) - \text{норм.}$$

Интуиция - близк. полож. отсюда

Нелинейный - осн. класс
 Регрессия - линейная \rightarrow линейн / линейн

STOLP: δ - порог фильтрации выбросов
 ℓ_0 - доп. доп. ошибки

- выкидыш выбросы ($M(x_i) < \delta \rightarrow \text{delete}$)
- $\Omega = \{ \text{argmax}_{x \in X^t} M(x, X^t) / y \in Y \} - \text{исл. син.}\}$ - нач. классов
- while $\Omega \neq X^t$:
 - $E := \{x_i \in X^t / \Omega : M(x_i, \Omega) < 0\}$
 - if $|E| < \ell_0$ exit
 - else $\Omega \leftarrow \Omega \cup \text{argmin}_{x \in E} M(x, \Omega)\}$

Не эффективно - можно забывать не исключать

⑥ Вероятн. постановка

$X \times Y$ — вероятн. пространство с
 $P(x, y) = P(y) p(x|y)$.

↑
↑
— априорные вер. на классы P_y

Задачи:

- X^t , $p(y|x) = P_y p_x(x)$. Построим элпс.
- если P_y и $\hat{p}_y(x)$ $\forall y \in Y$.
- по всем $P_y(x)$ и P_y построим элпс. $a(x)$ линии. б-ые элпс. классиф.

Решение

$$P(\mathcal{R}|y) = \int_{\mathcal{R}} p_y(x) dx; \quad \mathcal{R} \subset X$$

$a: X \rightarrow Y$ априорн. $A_y = \{x \in X / a(x) = y\}$
 $P_y P(A_s|y)$ — вер. отнесения x классу y и s
 $(y, s) \sim$ величина параметр λ_{ys} . (однако $\lambda_{yy} = 0$)
 $\lambda_{ys} > 0 \quad y \neq s$

Решение среднего риска:

$$R(a) = \sum_{y \in Y} \sum_{s \in S} \lambda_{ys} P_y P(A_s|y).$$

Если $\lambda_{ys} = \text{const}$ $\forall s \neq y$, то $R(a) = \text{вер. ошибки}$ на a

Оптим. базис. реш. правило:

$$a(x) = \arg \min_{s \in S} \sum_{y \in Y} \lambda_{ys} P_y p_y(x) - \text{минимум } R(a)$$

Если $\lambda_{yy} = 0$, $\lambda_{ys} = \lambda_y \quad \forall s, y \in Y$
 то

$$a(x) = \arg \max_{y \in Y} \lambda_y P_y p_y(x).$$

Разгл. поверхности между классами s, t —
 т.к. $x \in X$, что макс. $a(x)$ достигается
 одновременно при $y=s, y=t$

Аналогично $P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P_y(x) P_y}{\sum_{s \in S} P_s(x) P_s}$

$$\hat{P}_y(x) = \sum_{y \in Y} \lambda_y P(y|x)$$

$$\text{отсюда} \quad r(x) = \arg \max_{y \in Y} \lambda_y P(y|x)$$

Миним. $R(a)$, дает элпс. реш. правило,
 называемое базисовым риском
 Если $\lambda_y = 1$, то 'принцип максимума априор. вер.'

Послед.

$$X_y^t = \{(x_i, y_i)\}_{i=1}^t, \quad |y_i = y\} - \text{подвыборка превидений}$$

$$\hat{P}_y = \frac{\ell_y}{\ell}, \quad \ell_y = |X_y^t|, \quad y \in Y$$

— возвращение к P_y при $t \rightarrow \infty$

Построение: $X^m = \{x_1, \dots, x_m\}$ — выборки сущ.,
 и извлечение соотв. $p(x)$. Построим
 априорное правило отсюда — $\hat{p}(x)$,
 предикт. $p(x)$ на всем X .

Классификац. базис: пусть $x \in X$ определяется
 и числовыми признаками: $f_i: X \rightarrow \mathbb{R}$
 $x = (\xi_1, \dots, \xi_n)$, $\xi_i \in \mathbb{R} = f_i(x)$.

Если $P_y(x) = P_{y1}(\xi_1) \cdots P_{yn}(\xi_n)$, где $P_{yi}(\xi_j)$ —
 конст. распред. $j = 1 \dots n$
 Тогда $a(x) = \arg \max_{y \in Y} (\ln \lambda_y \hat{P}_y + \sum_{j=1}^n \ln \hat{P}_{yj}(\xi_j))$

⑦ Вер. постановка

В стат. постановке данные неприм. ($f_i \sim$ норм.),
 неполные (не все x)

Вер. постановка: $\exists p(x, y)$, выбираем ℓ пакет:

$$X^t = (x_i, y_i)_{i=1}^t, \quad y^*(x) \text{ можно превидеть}$$

и будем

$$P(y|x) = \delta(y - y^*(x)), \quad \delta(z) - \text{функция-дел}$$

$$P(x, y) = p(x) p(y|x)^{\delta}$$

Задача $\varphi(x, y, \theta)$ — выбора обес. плотности.

$L(\theta, X^e) = \prod_{i=1}^n \varphi(x_i; y_i; \theta)$. Чем θ раз L макс.
— групповое правдивое (принцип макс. правд.)

Вместо максимизации L удобно минимиз. — т.к.

$$-\ln L(\theta, X^e) = -\sum_{i=1}^n \ln \varphi(x_i; y_i; \theta) \rightarrow \min_{\theta}$$

Если в этом б. эмпир. рисце $\lambda = -\partial/\partial \theta \varphi(x, y, \theta)$,
то $Q(a, X^e) = L(\theta, X^e)$.

(8) Непараметр. оценка / классификация

• Непараметр. оценка плотности:

$$X - конечн., |X| \leq m. \hat{P}(x) = \frac{1}{m} \sum_{i=1}^m [x_i = x]$$

Непр. оценки \uparrow

$$\text{Непр. оценки: } p(x) = \lim_{h \rightarrow 0} \frac{1}{2h} P[x-h, x+h]$$

$$\hat{P}_h(x) = \frac{1}{2mh} \sum_{i=1}^m [x-x_i| < h] - непр. оценк.$$

$$\hat{P}_h(x) = \frac{1}{mh} \sum_{i=1}^m K\left(\frac{x-x_i}{h}\right) - локальная \underset{h}{\overset{\uparrow}{\text{оценка}}}$$

оценка Парзена-Розенблата

K -это $\int K(z) dz = 1$

Примеч. едро $K(z) = \frac{1}{2}[|z| < 1]$ ~ пристенка
точнее — диспр. оценки.

• X^m пристен., выбрана из плотности расп. $p(x)$
 $K(z)$ нефр., имеет шир. квадрат $\int K^2 dz < +\infty$
 $\lim_{m \rightarrow \infty} h_m = 0$, $\lim_{m \rightarrow \infty} nh_m = \infty$

Тогда $\hat{P}_{h_m}(x) \rightarrow p(x)$ при нормах всех x .

• Многомерный непр. оценки:

$$\hat{P}_h(x) = \frac{1}{m} \sum_{i=1}^m \prod_{j=1}^m \frac{1}{h_j} K\left(\frac{|f_j(x) - f_j(x_i)|}{h_j}\right)$$

• Граф. непр. н-бо: $\hat{P}_h(x) = \frac{1}{mV(h)} \sum_{i=1}^m K\left(\frac{\delta(x) - \delta(x_i)}{h}\right)$
нормированная интегр.

Задачи на разн. окна Δy :

$$\hat{P}_{y,h}(x) = \frac{1}{\epsilon_y V(h)} \sum_{i=1}^c [y_i = y] K\left(\frac{\delta(x, x_i)}{h}\right)$$

$$\text{Тогда } a(x, X^e, h) = \max_{y \in Y} \sum_{i=1}^c [y_i = y] K\left(\frac{\delta(x, x_i)}{h}\right)$$

Как подобрать параметры?

— Евро — не важно. Пусть выбрать члены.

— Окно — лучше делим. до k -го состояния

$$\text{тогда } Q(h, X^e) = \sum_{i=1}^c [a(x_i; X^e; h) \neq y_i] \rightarrow \min_h$$

Применим разномерные — если $\varphi(x, x')$ симметрия
на сумме координат и dim орен. больше,
объекты (вс.) могут быть оены различными.
Внешн. — помнить разн. ($\Sigma \nu_i / \nu_{\text{внешн.}}$ информативн.)

(9) Линейная клас. Граф. спос.

$$Y = \{-1, 1\}, T^e = \{t(x_i; y_i)\}_{i=1}^n$$

$$a(x, w) = \text{sign}(\langle w, x \rangle - w_0) = \begin{cases} \sum_{i=1}^n w_i f_i(x) - w_0 \\ \text{sign}(\langle w, x \rangle) \text{ если } \sum_{i=1}^n f_i = -1 \end{cases}$$

Модель класса: w_0 — порог активации.

Сумма зарядов функ. на соотв. весе w ,
плюс φ -е активации превращение
на $> < w_0$.

Реж. активации: хеббинг, симпл., макр. гамм.

$$a(x, T^e) = \delta\left(\sum_{i=1}^n w_i f_i(x) - w_0\right)$$

Граф. спос:

$$Q(w) = \left(\frac{1}{\epsilon}\right) \sum_{i=1}^c L(M_i; w) = \frac{1}{\epsilon} \sum_{i=1}^c L(\langle w, x_i \rangle \neq y_i) \rightarrow \min_w$$

$$w_0^* - перв. начальн. веса \quad w^{k+1} = w^k - \mu \nabla Q(w^k)$$

$$W - \text{нек. градиент}, \quad w = w - \eta Q'(w) \xrightarrow{\text{диф. } \frac{\partial L}{\partial w_i}} \left(\frac{\partial L}{\partial w_i} \right)_{i=1}^n$$

$$w^{k+1} = w^k - \eta \sum_{i=1}^n L'(\langle w, x_i \rangle, y_i)$$

неконвекс.

(10) Симпл. градиент + зврещение

w^0 - нач. ве.

$Q = \sum_i L(\langle w, x_i \rangle, y_i)$ - симпл. минимиз. оценка
 x_1, \dots, x_n - неконв. переход (расщеплен?)

$$\begin{aligned} w^{k+1} &= w^k - \eta \sum_i L'(\langle w^k, x_i \rangle, y_i) x_i \\ Q^{k+1} &= (1-\eta) Q^k + \eta \sum_i L'(\langle w^k, x_i \rangle, y_i) \end{aligned}$$

минимиз. оценка

Основополагающее, если

w симпл. или Q не линейны
 η -параметр оптимизации $\approx \frac{1}{C}$

Минимиз. w : $w_i = \text{random}(-\frac{1}{2n}, \frac{1}{2n})$.

Можно иначе. Можно: $w_j = \frac{\langle y_j, f_j \rangle}{\langle f_j, f_j \rangle}$

Плюсы: SG:

+ быстрота, отображение на линейные класс. и лог. регр.

+ демонстрирует обусловленность

+ для больших выборок

Недостатки:

- Q многостремителен \rightarrow можно застрять в лок. мин.,
- сходимость если n конечна или C мало.

- Демонстрация правила Гудисса $L(M) = (M-1)^2$
 Пусть $w = w - \eta (\langle w, x_i \rangle - y_i) x_i$
- Переопределение правила Хетда:
 $\begin{cases} \langle w, x_i \rangle > y_i \Rightarrow w := w + \eta x_i y_i - \text{const} \\ L = (-M)_+ \end{cases}$
 $\Rightarrow w = w + \eta (\text{sign}(\langle w, x_i \rangle) - y_i) x_i$
- Теор. Хетдса: где можно разделяющих T^+
 бе схорение и разделяем $\exists i, j > 0 (\langle \tilde{w}, x_i \rangle > y_i > \delta)$
 (с правильной Хетдой)
- Оптимизация:
 - Переопределение признаков $x_i^j = \frac{x_i^j - x_{\text{ср}}^j}{x_{\text{срвнр}}^j}$

- Переход - по очереди из разных классов;
- Год. весы - оптимизация с оценкой;
- Квадр. регуляризация (weights decay)
 $Q_\tau(w) = Q'(w) + \tau w$
- $Q_\tau = Q(w) + \frac{\tau}{2} \|w\|^2 \Rightarrow w = w(1-\eta\tau) - \eta Q'(w)$
 Нужно подобрать τ , но надо w всегда уменьшить.
- Весы шага - $\eta_t = 1/t$ - уменьшение с шагом.
 Итера. шаг можно аналогичным образом из $Q(w - \eta Q'(w)) \rightarrow \min$
- Выдавливание из лок. минимумов - в рамках
- Равнот. оценок - когда можно - можно винни. критерий наил. роста.

(11) Регуляризация - переход решения задачи обобщит.

- Квадрат. (см. 10) - опр. $\|w\|$, но забывает об ограничении роста. Поэтому, в частности, при линейном полиноме,
- В задачах вер. поискования есть принцип максим. совместного правдоподобия гауссовых моделей

$$L_g(w, X^t) = \underbrace{\ln p(X^t; w; g)}_{p(X^t; w) p(w; g)} = \underbrace{\sum \ln p(x_i, y_i | w)}_{\text{лок. правдоподобие}} + \underbrace{\ln p(w; g)}_{\text{регуляризатор}}$$

Плюсы:

- Квадратичный регуляризатор / квадратичный
- $\ln p(w, b) = \ln \left(\frac{1}{(2\pi)^n} \exp \left(- \frac{\|w\|^2}{2b} \right) \right) = -\frac{1}{2b} \|w\|^2 + \text{const}(w)$
- Практический
- $\ln p(w, C) = \ln \left(\frac{1}{(2C)^n} \exp \left(- \frac{\|w\|_2^2}{C} \right) \right) = -\frac{1}{C} \|w\|_2^2 + \text{const}(w)$

↑
 Чертежка $\|w\| = \sqrt{\sum |w_i|^2}$

Минимум правд.
 отбор признаков \rightarrow какие-то w_i будут 0

⑫ SVM.

$$X = \mathbb{R}^n, Y = \{-1, 1\}$$

$$\zeta(x) = \text{sign}(\langle w, x_i \rangle - w_0)$$

↑ определяет пороговый классиф.

График, видно что x^* является разделимым. γ

$$Q(w, w_0) = \sum [y_i(\langle w, x_i \rangle - w_0) \leq 0] = 0$$

где каких-то w, w_0

Как выбрать оптим. разделяющие плоскости?

- Компьютерка: $\min_{i=1..e} y_i (\langle w, x_i \rangle - w_0) = 1$
 $\{ -1 \leq \langle w, x_i \rangle - w_0 \leq 1 \}$ — ненес., разделя. классы

Максимизируем ширину:

$$\langle x_+ - x_-, \frac{w}{\|w\|} \rangle = \frac{2}{\|w\|} \rightarrow \begin{cases} \langle w, w \rangle \rightarrow \min \\ y_i (\langle w, x_i \rangle - w_0) \geq 1 \\ \forall i \in 1..k \end{cases}$$

Мин. разделимость

$$\boxed{\zeta_i \sim \text{одинакова для:}} \quad (*) \quad \begin{cases} \frac{1}{2} \langle w, w \rangle + C \sum \zeta_i \rightarrow \min \\ y_i (\langle w, x_i \rangle - w_0) \geq 1 - \zeta_i, \forall i \\ \zeta_i \geq 0 \end{cases}$$

Регуляризация

$M_i = y_i (\langle w, x_i \rangle - w_0)$. Выводим ζ_i через M_i :
 $\zeta_i \geq 0, \zeta_i \geq 1 - M_i$. В силу требований минимиз.
 $\sum \zeta_i$ какого-то гипот. максимиз. навязано.

$$\zeta_i = (1 - M_i)_+$$

$$\text{Тогда } (*) \text{ выглядит: } Q(w, w_0) = \sum_{i=1}^e (1 - M_i)(w, w_0)_+ + \frac{1}{2C} \|w\|^2 \rightarrow \min_w$$

норм. регуляриз.

Из графиков задача неразделяем.

$$(*) \quad \begin{cases} w = \sum_{i=1}^e \lambda_i y_i x_i & \text{— означает } w \text{ не м. } \lambda_i \\ \sum \lambda_i y_i = 0 \\ y_i + \lambda_i = C \quad i \in 1..e \end{cases}$$

$\sum \zeta_i$ — количество \vec{w}
 $\vec{x} \rightarrow k \vec{\zeta}$

Если $\lambda_i > 0$, то x_i — опорный вектор (support vec.)

- Периодичные — $\lambda_i = 0, y_i = C, \zeta_i = 0, M_i \geq 1$
- Опорные граничные — $0 < \lambda_i < C, 0 < y_i < C, \zeta_i > 0, M_i = 1$
 В этом случае на границе опорные
- $\lambda_i = C, y_i = 0, \zeta_i > 0, M_i < 1$ — парциальные

Тогда задача (глобальн.):

$$\begin{cases} -L(\lambda) = -\sum \lambda_i + \frac{1}{2} \sum \sum \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle \rightarrow \min \\ 0 \leq \lambda_i \leq C \\ \sum \lambda_i y_i = 0 \end{cases}$$

Когда решим эту задачу, w определяется через λ (на) а $w_0 = \min \{ \langle w, x_i \rangle - y_i : \lambda_i > 0, M_i = 1, i=1..e \}$

$$\text{Несколько } \alpha(x) = \text{sign} \left(\sum_{i=1}^e \lambda_i y_i \langle x_i, x \rangle - w_0 \right)$$

↑ не опорные векторы

- Или — если $\exists \psi / K(x, x') = \langle \psi(x), \psi(x') \rangle$

Теорема Марселя:

$$\begin{aligned} K-\text{ядро} \iff & \cdot K(x, x') = K(x', x) \\ & \cdot \int \int K(x, x') g(x) g(x') dx dx' \geq 0 \quad \forall g: x \rightarrow \mathbb{R} \end{aligned}$$

Ядра: $\langle x, x' \rangle^2, \langle x, x' \rangle^d, \delta(\langle x, x' \rangle), \exp(-\beta \|x - x'\|^2)$

Как использовать:

$$\alpha(x) = \text{sign} \left(\sum \lambda_i y_i K(x_i, x) - w_0 \right)$$

H — Суперлинейное пространство
 $\psi: X \rightarrow H \rightarrow$

⑬ Байесова классиф. и норм. диспр. анализ

Сам байесом — $P(R|y) = \int p(x, y) dx, \lambda y s > 0$
 или less

$$R(g) = \sum \sum \lambda y s P(A_s | y) - \text{сум. риск}$$

$$p(X, Y) = p(x) P(y|x) = P(y) p(x|y)$$

↑ независимость

— и означает в сумме

- Нормальный дистр. анализ:

Предполагаем, что $p_Y(x)$ $y \in Y$ - нормальное распределение $N(x, \mu, \Sigma) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu))$

\uparrow n -мерное норм. распред.

$$E\mathbf{x} = \mu, \quad E(x - \mu)(x - \mu)^T = \sum_{\text{матрица}}$$

$$\bullet \quad \sum = V S V^T, \quad (x - \mu)^T \sum^{-1} (x - \mu) = \\ (x - \mu)^T V S^{-1} V^T (x - \mu) = (x' - \mu')^T S^{-1}$$

Если признаки независимы, то $\sum = \text{diag}(6_1^2 \dots 6_n^2)$,
то линии уровня плоскостей —
эллипсоиды с центром μ и параллельны
координатам.

V - диагонализуем, потому что векторы оси параллельны осям.

Теорема: $\exists x \in X \mid \lambda_y P(y) p(x|y) = \lambda_s P(s) p(x|s)$

$$\bullet \quad \sum_{y+} = \sum_{y-} \rightarrow \begin{array}{l} \text{разр. поб-з квадратичн} \\ \text{линейн} \end{array} \quad \begin{array}{l} \text{байес. классификатор} \\ \text{задает поб-з} \end{array}$$

• Когда $\lambda_s P_s = \lambda_t P_t$, то $\sum_s = \sum_t = \sigma^2 I_n$, классы неизотропные и имеют одинак. форму.

Разр. н-т - первенчн. \varnothing, O_2, O_i - члены симметрии

Если не равнозначн. ($P_s \neq \lambda_t P_t$), то классы ближе к менее значимому классу

Число классов $> 2 \Rightarrow$ квадр. фундамент. иные кусочно-линейные

• Классы равнозначн. и равнозначн., тогда разр. поб-з:

$$\|x - \mu_s\|_\Sigma = \|x - \mu_t\|_\Sigma, \text{ т.е. } \|u - v\|_\Sigma = \sqrt{(u - v)^T \Sigma^{-1} (u - v)}$$

$\|u - v\|$ - расстояние Махаланобиса.

Макс. возможное правдоподобие:

$$L(\theta, x^m, w^m) = \sum_{i=1}^m w_i \ln \varphi(x_i, \theta) \rightarrow \max_{\theta} \quad W_i = \sum_{y: y_i=y} w_i$$

$$W^m = \{w_1, \dots, w_m\} \quad \hat{\mu}_y = \frac{1}{W_y} \sum_{i: y_i=y} w_i x_i$$

$$\hat{E}_y = \frac{1}{W_y} \sum_{i: y_i=y} w_i (x - \hat{\mu}_y) / (x - \hat{\mu}_y)^T$$

$$a(x) = \arg \max_{y \in Y} \left(\ln \lambda_y P(y) - \frac{1}{2} (x - \hat{\mu}_y)^T \sum_y^{-1} (x - \hat{\mu}_y) - \frac{1}{2} \ln \det \Sigma_y \right)$$

\uparrow байесовский норм. классификатор

- $\lambda_y < n \Rightarrow$ вырожденность Σ_y
- проблема избыточности параметров

Линейный дискриминант Рамера

Пусть все $\Sigma_i = \Sigma$. Тогда

$$\hat{\Sigma} = \frac{1}{\sum_{y \in Y} \# y_i=y} \sum_{i: y_i=y} (x_i - \hat{\mu}_{y_i}) (x_i - \hat{\mu}_{y_i})^T$$

или

$$a(x) = \arg \max_{y \in Y} (\lambda_y P(y) p(x|y)) =$$

$$= \arg \max_{y \in Y} \underbrace{(\ln(\lambda_y P(y)) - \frac{1}{2} \hat{\mu}_y^T \sum_y^{-1} \hat{\mu}_y)}_{\beta_y} + x^T \sum_y^{-1} \hat{\mu}_y$$

$$= \arg \max_{y \in Y} (\beta_y + x^T \alpha_y) = \text{sign}(\langle x, w \rangle - w_0)$$

Линейный дистр. Рамера — более отдален от байесовскому, но более устойчив чем квадр.

(если члены классов ~ нормальные)

Вероятность по ошибки $R(a) = \Phi(-\frac{1}{2} \|\mu_1 - \mu_2\|_\Sigma^2)$

$$\Phi(p) = N(x, \mu, 1)$$

(14) Логист. регрессия и ROC

Гипотеза: $a(x) = \operatorname{argmax}_{y \in Y} \lambda_y P(y) P(x|y)$

Линейный:

$$a(x, \theta) = \operatorname{sign}(\langle w, x \rangle)$$

Максимизация: $Q(\theta_0, T^L) = \frac{1}{L} \sum_i^L L(\theta_0, x_i) =$

$$= -\sum_i \ln P(x_i; y_i; \theta) \rightarrow \min$$

Коэффициенты в лог. дисперсии, анализ показывает, что коэффициент + правильный класс \Rightarrow о.н. байес. класс. линеек

$$P(x) = \exp(c(\theta) \langle \theta, x \rangle + b(\theta, \theta) + d(x, \theta))$$

тогда P — экспонента
Лог. класс., правильный класс — экспонента

Пусть $Y = \{-1, 1\}$, $X = \mathbb{R}^n$

$$a(x) = \operatorname{sign}(\lambda_+ P(+1/x) - \lambda_- P(-1/x)) = \operatorname{sign}\left(\frac{P(+1/x)}{P(-1/x)} - \frac{\lambda_-}{\lambda_+}\right)$$

С базисе для нашего Y

Теорема. логистическое правдоподобие $P(x|y)$ — экспон., членом правильные d, d , но о.н. в θ_y , тогда

$$1. a(x) = \operatorname{sign}(\langle w, x \rangle - w_0), w_0 = \ln(\lambda_+ / \lambda_-), w \text{ не з.бр. о.н. } \lambda_-, \lambda_+$$

$$2. P(y|x) = \tilde{b}(\langle w, x \rangle; y), \text{ где } \tilde{b}(z) = \frac{1}{1+e^{-z}} \text{ сигмоид}$$

Макс. правдоподобие $L(w, x^L) = \log \prod p(x_i; y_i) \rightarrow \max$, но: несправедл.

$$L = \sum_i \log \tilde{b}(\langle w, x_i \rangle; y_i) + \text{const}(w) \rightarrow \max$$

Много определений:

$$Q = \sum_{i=1}^L \log \left(1 + \exp(-\langle w, x_i \rangle; y_i) \right) \rightarrow \min$$

$$\Delta = \log \left(\frac{1}{1 + e^{-M}} \right) \text{ — логарифм. } \varphi \rightarrow 1 \text{ нонлип.}$$

Прав. сигнал: $\tilde{b}'(z) = \tilde{b}(z) \tilde{b}'(-z)$

тогда $w = w + \eta \sum_i x_i \tilde{b}'(-\langle w, x_i \rangle; y_i)$

Аналогично правилом Хадда:

$$w = w + \eta \sum_i x_i \underbrace{[\tilde{b}(w, x_i)]}_{\text{если } \langle w, x_i \rangle < 0} \frac{\tilde{b}'(w, x_i)}{\tilde{b}'(w, x_i)}$$

• Метрики

$$\text{Sens} = \text{Recall} = \frac{TP}{P} \quad \text{Precision} = \frac{TP}{TP + FP}$$

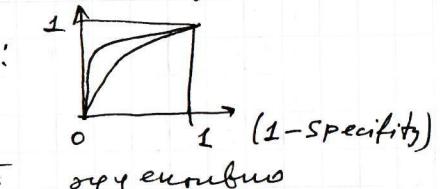
$$\text{Accuracy} = \frac{TP + TN}{P + N}$$

$$\text{Specificity} = \frac{TN}{N}$$

Sensitivity

Сумма метрик, которая считает AUC

$$F_B = (1+\beta^2) \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$



(15) Нер. предикторы

Концепт — $\varphi: X \rightarrow \{0, 1\}^Y$. φ покрывает X если $\varphi(x) = 1$

Правило — правило, покрывающее много областей единого класса и мало — других

φ многофункциональна для \leftarrow класс

$$p(\varphi) = |\{x_i | \varphi(x_i) = 1, y_i = c\}| \rightarrow \max$$

$$n(\varphi) = -|\{x_i | \varphi(x_i) \neq c\}| \rightarrow \min$$

$p \rightarrow$ true positive, $n \rightarrow$ False positive

$$\text{Precision} = \frac{p}{p+n} \rightarrow \max$$

$$\text{Accuracy} = \frac{p-n}{p-n} \rightarrow \max$$

$$\text{Linear cost acc.: } p - C_n \rightarrow \max$$

$$\text{Relative acc. : } \frac{p}{p} - \frac{n}{N} \rightarrow \max$$

• Правила/критерии: многофункциональны.

$$\bullet \varepsilon\text{-}\delta \text{ правило: } E_c(\varphi, T^L) = \frac{n_c(\varphi)}{p_c(\varphi) + n_c(\varphi)}, D_c(\varphi) = \frac{p_c(\varphi)}{L}$$

$\varphi(x) = \varepsilon\text{-правило}, \text{ если } \frac{E_c(\varphi, T^L)}{D_c(\varphi, T^L)} \leq \varepsilon \text{ и } \frac{E_c(\varphi, T^L)}{D_c(\varphi, T^L)} \geq \delta$

$n_c(\varphi) = 0 \Rightarrow$ правило погреш.

$$\bullet \text{Сумм. правило: } U_{PN}(p, n) = \frac{C_p^p C_n^n}{C_{p+n}^{p+n}}$$

Полинома (comprehension) φ на T^e

$$I_c(\varphi, T^e) = -\ln H_{PN_c}(P_{c(0)}, n_c(\varphi))$$

φ — статистическое правило, если

$$I_c(\varphi, T^e) \geq d \quad \text{с геом. базисами } d \\ \text{(помимо нее Ренера)}$$

• Entropy-based правило

$$H(q, p) = -q \log q - p \log p$$

$$\hat{H}(P, N) = H\left(\frac{P}{P+N}, \frac{N}{P+N}\right) \text{ — энтропия Бенфорда}$$

$$\hat{H}_\varphi(P, N, p, n) = \frac{p+n}{P+N} \hat{H}(P, N) + \frac{P+N-p-n}{P+N} \hat{H}(P-p, N-n)$$

$$IGain_\varphi(\varphi, T^e) = \hat{H}(P, N) - \hat{H}_\varphi(P, N, p, n)$$

$$\varphi \text{ — entropy-based, если } IGain_\varphi(\varphi, T^e) \geq 6_0$$

• Частичные правила

$$R(x) = \bigwedge_{i \in J} [a_i \leq f_i(x) \leq b_i]$$

$$R(x) = \left[\sum_{j \in J} [a_j \leq f_j(x) \leq b_j] \geq d \right]$$

$$\text{— полносность: } R = [\sum w_j f_j(x) \geq w_0]$$

$$\text{— мер } R = [z(x, x_0) \leq r_0]$$

• Правил замены.

Алгоритм эвакуации — на каждом шаге лупоровение правил, выбираем лучше, лучше — самое информативное согласно

16) Деревья решений

• Коды — бинарные представления. Множество — классы для значений (гене. дерево)

Пример дерева с извлечением имен:

• Характ. альт. постр. дерева: 1D3

<u>1</u>	<u>0</u>	<u>2</u>
<u>0</u>	<u>1</u>	

Learn ($U \subseteq X^e$):

если $\forall x \in U$ из класса $C \Rightarrow$ верн. есть с

$B = \arg \max_{B \in \mathcal{B}} I(B, U)$ — правило с max информации

$U = U_1 \cup U_2$ — разделило по B $\uparrow I - IGain$

U_1 или $U_2 = \emptyset \Rightarrow$ верн. нет

шаги рекурсивно строим дерево

(если задача стоял — примеряет, то

б. случае если он верен — возвр.
Более популярн. классы из $\{B, B_2\}$)

B — семейство правил. В простом случае $f_i(x) > m_i$ или $f_i(x) = d_i$ или подмножеств

1D3:

+ Частичнорешающие

+ Простота, линейн.

+ Разноголосные данные

+ Не требует отдельн. отклас.

(можно в легког. данных откласы)

- переобучение за счет

автоматич. дерева

- высокая чувств. к

шуму

(мон. — примеряет: • один из классов лучше

$$I(B, U) < threshold$$

$$|U| < threshold$$

• tree height $> h$

• C4.5 Pruning

$\forall V \in V$ выбир.:

S_V — подмножество X^k , содержащих до V

$S_V = \emptyset \Rightarrow$ вернутие нет V , c_V — Маж. класс (U)

проверим на кол-во ошибок при классиф. S_V : V , невып/правиль. сначала,

V с другим классом C'

Заменим на v_0 , где ошибки меньше.

$X^k -$ ~~подмножество~~. Быстро

(17) Методы линейн. регрессии

$$Q(\alpha, X^e) = \sum_{i=1}^n (g(x_i, \alpha) - y_i)^2$$

$\alpha^* = \underset{\alpha \in \mathbb{R}^p}{\operatorname{argmin}} Q(\alpha, X^e)$ — это задача можно решать через миним. квадр.

Задачами решения $g(x, y)$ называются:

- Квадратичные — Весом.
- $g(x, \alpha) = \alpha \cdot x$. Весом или $\alpha(x) = \alpha \cdot x$ — линейный минимум. квадратов

$$Q(\alpha, X^e) = \sum_{i=1}^n w_i(x_i)(\alpha - y_i)^2 \rightarrow \min_{\alpha \in \mathbb{R}^p}$$

Лучше $w_i(x_i) = K(g(x_i, x_i)/h)$, $K: [0, \infty) \rightarrow [0, \infty]$

Приравнив $\frac{\partial Q}{\partial \alpha} = 0$, получим: (одностр. мин. H-B)

$$\alpha_h(x, X^e) = \frac{\sum_{i=1}^n y_i K(\frac{g(x_i, x_i)}{h})}{\sum_{i=1}^n K(\frac{g(x_i, x_i)}{h})}$$

Теорема: Хорошее близкое, хорошее близко (ограничено), $E(y^2/x) < \infty$, $h \rightarrow 0$ и $h \rightarrow \infty$

Тогда $\alpha_h(x, X^e) \rightarrow E(y/x)$ в нормах

- Выбор близко — разница между квадратом остатков и квадратом остатков может быть оптимальной, амортизацией, передаваясь постепенно близким соседям.
- Выбор оптим. — лучше динамически до k-го соседа. Но 100 подборов. $(\sum (a_h(x_i, X^e) / h - y_i)^2 \rightarrow \min_h)$
- LOWESS — локально убывающее сглаживание

Введен $\hat{y}_i = \tilde{K}(\epsilon_i)$, где $\epsilon_i = |a_h(x_i; X^e / h - y_i)|$

Алгоритм:

$$\left| \begin{array}{l} a_i = a_h(x_i; X^e / h - y_i) = \frac{\sum_{j=1, j \neq i}^n y_j w_j K(\frac{g(x_j, x_i)}{h})}{\sum_{j=1}^n w_j K(\frac{g(x_j, x_i)}{h})} \\ w_i = \tilde{K}(|a_i - y_i|) \quad k_{i-1..l} \\ \text{until } \{w_i\} \text{ стабилизируется} \end{array} \right.$$

Минимизование уравнений и задача разделяется.

$$\tilde{K}(\epsilon) : \text{техническая функция } \tilde{K}(\epsilon) = [\epsilon \leq \epsilon^{(r-t)}]$$

механизм: $\tilde{K}(\epsilon) = K_Q\left(\frac{\epsilon}{\epsilon_{\text{медианы}}}\right)$

↑
медиана $\epsilon_{\text{медианы}}$, $\epsilon^1 \leq \epsilon^2 \leq \dots \leq \epsilon^r$

(18) Линейная регрессия / SVD

$$g(x, \alpha) = \sum_{i=1}^n \alpha_i f_i(x). \quad y = (y_i)_{i=1}^n - \text{левый вектор} \\ \alpha = (a_i)_{i=1}^n - \text{вектор норм. признаков}$$

$$Q(\alpha) = \|F\alpha - y\|^2 \quad F - l \times n \text{ матр. признаков}$$

Недостаток минимума $\Rightarrow F^T F \alpha = F^T y$
y есть сист. лин. уравн., если $F^T F$ н.н. невыполн.

$$\alpha^* = (F^T F)^{-1} F^T y = F^+ y. \quad F^+ - \text{пseudooобратная}$$

$$Q(\alpha^*) = \|P_F y - y\|^2 \quad P_F = \underbrace{F F^+}_{\text{расширенный}} \text{ — проекционное}$$

Квадрат гипотезы паритета с y на $\text{col}(F)$

• SVD $l \times n$ матрица ранга n — $F = V D U^T$
на 1. D — diag($\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}$) — λ_i — собственные значения FFT и $F^T F$

на 2. V ортогональна, $V V^T = I$, v_i — ортог. вект. FFT
на 3. U ортог., $U U^T = I$, $-U^T$ — ортог. вект. $F^T F$

$$F^+ = U D^{-1} V^T = \sum_{j=1}^n \frac{1}{\sqrt{\lambda_j}} v_i v_i^T; \alpha^* = F^+ y = \sum_{j=1}^n \frac{1}{\sqrt{\lambda_j}} v_i (v_i^T y)$$

$$F \alpha^* = \sum v_i (v_i^T y) - \text{МНК-аппрок. y}$$

• Если $\mathcal{E} = F^T F$ имеет ненул. ранг, то ее псевдо обратима. Для решения оптимизационных задач использует L2 норма или регуляризацию.

• Мультиколлинеар. — имеет полн. ранг, но много обусловлено ($\lambda_{\max}/\lambda_{\min} \geq 10^2 \sim 10^4$)

• Ridge / гребневая регрессия

$$Q_T(\alpha) = \|F\alpha - y\|^2 + T\|\alpha\|^2. \quad \text{В сущ. мультиколлинеар. сист. много d дают мин. } \|T\| \text{ называют выбросом}$$

$$\frac{\partial Q(\alpha)}{\partial \alpha} = 0 \Rightarrow \alpha^* = (F^T F + \tau I_n)^{-1} F^T y$$

Все собств. значения убес. на τ , а собств. вектора не изменились. Матрица собств. векторов обозн.

$$\alpha^* = (UD^2U^T + \tau I_n)^{-1} UDV^T y = \sum_{j=1}^n \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} v_j (v_j^T y)$$

- Лasso - неприм. регуляризатор.

$$\begin{cases} Q(\alpha) = \|F\alpha - y\|^2 \rightarrow \min_{\alpha} \\ \sum_{j=1}^n |\alpha_j| \leq \beta \end{cases} \leftarrow \text{наприм. регуляризация}$$

(19) Нелинейные методы линеар. регрессии

$$Q = \sum_{i=1}^e (f(x_i, \alpha) - y_i)^2 \quad \left\{ \begin{array}{l} \text{лучш. приг. спосб!} \\ \text{помимо, это супр. минимум, не лин.} \\ \text{лобс - можно в рабочих} \end{array} \right.$$

- Нелинейный - Равен

$$\alpha^{t+1} = \alpha^t - h_t (Q''/\alpha^t)^{-1} (Q^{t+1}/\alpha^t)$$

\uparrow \uparrow
наш(1) расчет $\delta \alpha^t$ \uparrow градиент $Q \delta \alpha^t$

График и расчет вспом. численно. Можно - обратиться Q'' на некор. итерации

Если $f \in C^2(\mathbb{R})$ - функ. диф. дважды, то линеаризован:

$$f(r_i, \alpha) = f(x_i, \alpha^t) + \sum_{j=1}^p \frac{\partial f}{\partial \alpha_j}(x_i, \alpha^t)(\alpha_j - \alpha_j^t)$$

тогда в линейное приводит к лин. прям.

и формула:

$$\alpha^{t+1} = \alpha^t - h_t (F_t^T F_t)^{-1} F_t^T (f^t - y)$$

δ \rightarrow $\|F_t \delta - (f^t - y)\|^2 \rightarrow \min$

F_t - лин. прям.
 $f^t = f(x_i, \alpha^t)_{i=1}^n$

- Обобщ. лин. методы

$f(x, \alpha)$ линейна, $g(f)$ - нем.

$$Q(\alpha, X^t) = \sum_i (g(\sum_j \alpha_j f_j(x_i)) - y_i)^2 \rightarrow \min$$

$g(f)$ - заданная, напр. дифференцируемая. $\exists \alpha$ - реш.

Линеаризован $g(z)$ в окр. какого ли z_i .

$$g(z) = g(z_i) + g'(z_i)(z - z_i)$$

Тогда Q аппроксимируется \hat{Q} :

$$\begin{aligned} \hat{Q}(\alpha, X^t) &= \sum (g(z_i) + g'(z_i)(\sum_{j=1}^n \alpha_j f_j(x_i) - z_i) - y_i)^2 = \\ &= \sum_{i=1}^e \underbrace{(g'(z_i))^2}_{w_i} \left(\sum_{j=1}^n \alpha_j f_j(x_i) - \underbrace{z_i - \frac{y_i - g(z_i)}{g'(z_i)}}_{\tilde{f}_i} \right)^2 \end{aligned}$$

Получаем
линейную
матричную
задачу
решение
этой
задачи
за
помощью
прямых
методов.

(20)

(26) Кейрс, правило Хорда..

Розенблatt: $a_w(x, T^k) = \delta \left(\sum_{i=1}^n w_i x^{(i)} - w_0 \right)$

↑ хобусаң!

Мак McLulch-Pitts: $a_w(x, T^k) = \delta \left(\sum w_i f_i(x) - w_0 \right)$

Рукавин актөзесін: $(1 + e^{-z})^{-1}$ – синусоид.
 $(2\delta/(2z) - 1) = \text{th}(z)$

Правило Розенблата: $Y = \{0, 1\}$,
 $w^{k+1} = w^k - \eta (a_w(x_k) - y_k)$

Правило Хорда:
 $w^{k+1} = w^k + \eta x_k y_k$ [$\langle w^k, x_k \rangle y_k < 0$]

Лемма – правило: $L = (\langle w, r \rangle - 1)^2$
 $w^{k+1} = w^k - \eta (\langle w, r \rangle - y_k)$

• Проблема полиномі

or, and, not – выражения. \oplus – через композицию
 $x^1 \oplus x^2 = [(x^1 \vee x^2) - (x^1 \wedge x^2)] \cdot \frac{1}{2}$

Любые bool φ-ыи выражения ($\oplus \wedge \Phi$)

Теорема (Горданс): X компактно, $C(X)$ – амалық
 шергүйінде Φ – X , F -ны. көпжет, замыктуғас олар.
 көпжеттік шергүй. Φ , сәттілікке 1. Тогда F толе
 $\Phi \in C(X)$

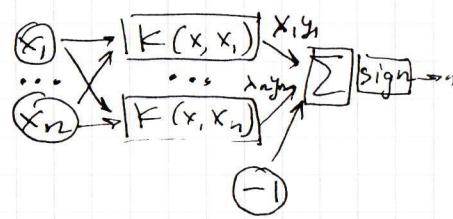
Кароғе шергүй. сон шартта в шергүй Φ -ны.

• Многомодально: – распределительность

1. Входной шаг: x_1, \dots, x_n , $-1 \leftarrow$ всегда есть в
 2. Сортировка (ascension)

3. инициализация $\sum - b_n \leftarrow$

SVM на шергүй. сенсіс:
 (с логарифм) b – опорн. вектор



(27) Обратное проекция

Множес. сенсіс, $Y \in \mathbb{R}^M$ $X \in \mathbb{R}^n$

Вынуждене значение на обратные сенсіс: x_i :

$$a^m(x_i) = \delta_m \left(\sum_{h=1}^H w_{ih} u^h(x_i) \right), u^h(x_i) = b_h / \sum_{i=1}^M w_{ih} \sigma^h(x_i)$$

↑ 2 сенсіс
↑ 1 сенсіс

$$\frac{\partial Q(w)}{\partial a^m} = a^m(x_i) - y_i^m = \epsilon_i^m \quad Q(w) = \frac{1}{2} \left(\sum (a^m(x_i) - y_i^m)^2 \right)$$

$$\frac{\partial Q(w)}{\partial b_h} = \sum (a^m(x_i) - y_i^m) \delta_m' w_{ih} = \sum \epsilon_i^m \delta_m' w_{ih} = \epsilon_i^h$$

w_{ih} $\epsilon_i^h \delta_m'$
 $\epsilon_i^h \leftarrow \sum \epsilon_i^m \delta_m'$

Алгоритм:
 init $w_{jh} = \text{random}(-\frac{1}{2n}, \frac{1}{2n})$, $w_{hm} = \dots (\frac{1}{2n}, \frac{1}{2n})$
 прием \times оғ:

Выбраныи предмет $(x_i, y_i) \in X^e$ сенсіс

Поставлены u_i^h, a_i^m
 $\epsilon_i^m = a_i^m - y_i^m$, $Q_i = \sum_{m=1}^M (\epsilon_i^m)^2$

обратный оғ:

$$\epsilon_i^h = \sum \epsilon_i^m \delta_m' w_{ih} \quad \forall h \in 1 \dots H$$

изменение. оғар

$$w_{hm} = w_{hm} - \eta \epsilon_i^m \delta_m' i^m$$

$$w_{jh} = w_{jh} - \eta \epsilon_i^h \delta_h' x_i^j$$

$$Q = \frac{(e-1)}{e} Q + \frac{1}{e} Q_i$$

Одним:

- + диффузивность $O(Mn + HM)$ – именеме сортировка, пац., пересечение, доказательство непротиворечия...

Нормати:

1. Инициализация w_0 бары (init) в промежах $[-\frac{1}{2n}, \frac{1}{2n}]$
2. Спосабы, градиент, сенсіс классификациянын виборы

3. Множества сопряженных градиентов — алгоритм.
нед. батч.

• Оптимизация структур:

1. Число слоев — $\{1..3\}$
2. Выбор кол-ва нейронов в каждом слое
визуально или через CV или H
3. Динам. добавл. нейронов при росте ошибки
4. Optimal brain damage — прореживание слоев
таким нейронами, которые не влияют на Q

(28) Композиции, бусинки

• n алгоритмов с р. классификации (алгоритм) —
 $P_1 \dots P_n \approx P$, независимые.

$$Pr_{tot} = P^n + nP^{n-1}(1-P) + \dots + C_n^{n/2} P^{n/2} (1-P)^{n/2}.$$

Задача: $H = \{h(x, a) : X \rightarrow R \mid a \in A\}$ $R = \text{Rum}, R''$
Найти наиболее точный
алгоритм в H.

$h_1 \dots h_N : X \rightarrow R$ — базовые алгоритмы

$$H_f(x) = C(F(h_1(x) \dots h_r(x)))$$
 — композиция
 ↓
 коррекц. операции
 правило выбора

В задачах классиф. $f(x) = F(h_1(x) \dots h_r(x)) = \sum d_f h_i(x)$
называется базис. гипотезами.

В регрессии правило $C = id$, называю то Y близко
В классиф. на $\{-1, 1\}$ $C = sign$

Бусинка $Y = h-1, 1\}$, $C = sign$, базоб. ам. булев. $\{-1, 0, 1\}$

$$\left. \begin{aligned} a(x) &= C(F(h_1 \dots h_r)) = sign\left(\sum d_f h_i(x)\right) \\ Q_f &= \sum_i^c [y_i \sum_{t=1}^T d_t \beta_t(x_i) < 0] \end{aligned} \right\}$$

Бусинка — это алгоритм добавления алгоритмов
или, то есть следующий сопряженный слой формируется
предыдущего

Пример: $MSE = \frac{1}{c} \sum_i^c (a(x_i) - y_i)^2$

Обычно $\ell_1(x) = \arg \min \frac{1}{c} \sum_i^c (\delta(x_i) - y_i)^2$

$$\ell_2(x) = \arg \min \frac{1}{c} \sum_i^c (\delta(x_i) + \delta(x_i) - y_i)^2$$

$$\ell_N(x) = \arg \min \frac{1}{c} \sum_i^c \left(\delta(x_i) - \left(y_i - \sum_{j=1}^{N-1} \delta_j(x_i) \right) \right)^2$$

(29) Градиенты. Бусинки

$$a_N(x) = \sum_i^N \ell_i(x) \quad L(y, z) = (y - z)^2 \\ = \log(1 + \exp(-yz))$$

Инициализация $\ell_0 = 0$
 $\ell_0 = \frac{1}{c} \sum_{i=1}^c y_i$ или сам. гаусс. y_i

Обычно $Q = \sum_{i=1}^c L(y_i, a_{N-1}(x_i) + \delta(x_i)) \rightarrow \min$

Для нахождения наимен. мин. вект. субградиент

$$s = -\nabla F, \text{ где } F = \sum_i^c L(y_i, a_{N-1}(x_i) + \delta_i)$$

Линейн. градиент. субградиент
мода $\ell_N(x) = \arg \min \frac{1}{c} \sum_{i=1}^c (\delta(x_i) - s_i)^2$

квадр. 4-я л
линейн., реальная
L-б. градиент

• Сокращение размера шага: $a_N(x) = a_{N-1}(x) + \gamma \ell_N(x)$
меньше $\gamma \rightarrow$ константа, но медленнее

• $\gamma \in N$ (много шагов) — подходит по CV
(гомо, фикс. одно.)

• Болтун (сплошн. слой)
относится к небольшим слоям

• Результат — сплошной плавающие деревья реш.
Решение $a = a^*$
Классификация: $\{-1, 1\}$. L-логистическая

$$P(\gamma=1|x) = \frac{1}{1 + \exp(-a^*(x))} \quad P(\gamma=-1|x) = \frac{1 - 1}{1 + \exp(-a^*(x))}$$

④) Фунд. алгоритм AdaBoost

$$a(x) = \text{sign}(F(b_1 + b_2 + \dots + b_l)), \quad a(x) = \text{sign}(\sum a_i b_i(x))$$

$$Q_T = \sum [y_i \sum a_i b_i(x_i) < 0] - \text{конт. боя симбок}$$

- Нбр 1: при добавлении b_t можно только один (пред. член)
- Нбр 2: пороговая Q_T заменяется наименьшими, чтобы

$$Q_T \leq \tilde{Q}_T = \sum_{i=1}^l \exp(-y_i \sum_{j=1}^l a_j b_j(x_i)) =$$

$$\sum \underbrace{\exp(-y_i \sum_{j=1}^l a_j b_j(x_i))}_{\text{exp}(-y_i \sum_{j=1}^l a_j b_j(x_i)) \cdot \exp(-y_i b_T(x_i))} \cdot \exp(-y_i b_T(x_i))$$

$$\tilde{W}^t = (\tilde{w}_1, \dots, \tilde{w}_l), \quad w_i = \tilde{w}_i / \sum_{j=1}^l \tilde{w}_j$$

$$u^t = (u_1, \dots, u_l) - \text{вектор весов}, \quad \sum u_i = 1$$

$$N(b, u^t) = \sum_{i=1}^l u_i [b(x_i) = -y_i]$$

$$P(b, u^t) = \sum_{i=1}^l u_i [b(x_i) = y_i]$$

1-N-P-вес
отказов от
классификации

все правильные
все ошибочные

Теорема 1. $\exists \forall u^t$ существует b_t , такое что $P(b, u^t) > N(b, u^t)$

может минимизировать Q_T если при

$$b_t = \arg \max_b [\sqrt{P(b, \tilde{W}^t)} - \sqrt{N(b, \tilde{W}^t)}]$$

$$\alpha_t = \frac{1}{2} \ln \frac{P(b_t, \tilde{W}^t)}{N(b_t, \tilde{W}^t)}$$

Теорема 2

На каком числе b_t : $\sqrt{P(b_t, \tilde{W}^t)} - \sqrt{N(b_t, \tilde{W}^t)} = p_t > 0$

при некотором $p_t > 0 \Rightarrow a(x)$ будет неправильной

за конечное число шагов

• Теорема 3 (основная)

Если для алг. не омн. от классификации и $\sqrt{W^t} \geq b_t \mid N(b, W^t) < \frac{1}{2}$, то минимум Q_T достигается при:

$$b_t = \arg \min_b N(b, \tilde{W}^t)$$

$$\alpha_t = \frac{1}{2} \ln \frac{1 - N(b_t, \tilde{W}^t)}{N(b_t, \tilde{W}^t)}$$

Алгоритм: X^l, Y^l - выборка. T - макс. число базовых

$w_i = 1/l$ - нормализация

for $t = 1..T$ пока не остановлено: (или \tilde{Q} не сокр.)

$b_t = \arg \min_b \dots$ из теоремы 3

$a_t = \frac{1}{2} \ln \dots$

$w_i = w_i \exp(-d + y_i b_t(x_i)) \quad i = 1..l \leftarrow$ пересчет

нормализовать w_i

• Summary

- + Хороший обогащ. способность
- + простота реализации
- + возможно снизить перекрестную проверку (если w_i - масштабированные)
- переобучение на шуме
- узконаправленный обогащ. Выборка
- неинвариантен к перестановкам

③) • Прогр. параметров

Большинство алгоритмов не умеет работать с бесконечными объектами

\Rightarrow идея - упростить один алгоритм модели на разных языках и синтаксисах

Subsampling - $k < n$ - подвыборка

Bagging - n раз берут из 1 эл-ты с возвратом.

BSM (random subspace method) - не изучался. Факт

Filtering -

1. Ищем b_1 на X_1 - первые m_1 объектов X^l
2. $\nexists m_2$ наз. правило:
 - добавл. в X_2 второе класс. объект
 - или первое
 Ищем b_2 на X_2
3. Добавление $\nexists m_3$ объектов, так что $b_1(x) \neq b_2(x)$

• Random forest.

Лес переобучается.

$$a(x) = \text{sign} \frac{1}{N} \sum_{n=1..N} b_n(x)$$

Алгоритм:

1. N случайных небольших $X_n, n = 1..N$

2. Соревнование

- поиск b такого чтобы не оказалось одинаковых

- процесс назначение responsibility - на основе признака ищется среди всех признаков разные g_{ij} .

- Признаки вектор. назначения для какой вершины

3. Обобщение:

$$a(x) = \frac{1}{N} \sum_{i=1}^N g_{ni}(x) - \text{нагр.}$$

$$a(x) = \operatorname{sign} \frac{1}{N} \sum_{i=1}^N b_n(x) - \text{нагр.}$$

Можно параллелизм.

(32) Синтез

Приму классы, назовем оние распредел.

$$p(x) = \sum_{j=1}^k w_j p_j(x), \quad \sum w_i = 1 \quad w_i \geq 0$$

$p_j(x)$ - функция правдоподобия
 w_j - ее априорная вероятность

$p_j(x) = \varphi(x; \theta_j)$ - параметр. симметрии

Выберем эпсилон из синтеза - выбир. $w_i \gg \log p_i(x)$

Задана норма k

Задана разбиение \mathcal{X}^m расп. по $p(x)$
оноим $\Theta = (w_1, \dots, w_k, \theta_1, \dots, \theta_m)$ зове Ψ, k

(33) EM-алгоритм expectation-minimization

Вектор G -вектор скрытых параметров.

• Может быть выполнено по Θ

• Красиво управ. разбивкой максимумом правдоподобия.

Алгоритм:

нека G, Θ не стабильны:

$$G \leftarrow \text{estep}(\Theta)$$

$$\Theta \leftarrow \text{mstep}(G, \Theta)$$

• Estep. $P(x; \theta_j)$ - логарифм. тоо, тоо x попадает в j -ий компонент синтеза

$$P(x; \theta_j) = P(x) \underbrace{P(\theta_j | x)}_{g_{ij}} = w_j p_j(x)$$

$G = \|g_{ij}\|$. Заметим, тоо $\sum_{i=1}^k g_{ij} = 1 \forall i$.

Очевидно

$$g_{ij} = \frac{w_j p_j(x_i)}{\sum_{s=1}^k w_s p_s(x_i)} \quad \forall i$$

но оно же бывает

$$\begin{aligned} \text{M-step} \text{ сводится к } & \left\{ \begin{array}{l} \theta_j = \operatorname{argmax}_{\theta_j} \sum_{i=1}^m g_{ij} \ln p(x_i; \theta) \\ w_j = \frac{1}{m} \sum_{i=1}^m g_{ij} \end{array} \right. \end{aligned}$$

• Критерий остановки: $\text{зубчик } \max |g_{ij} - g_{ij}^0| > \delta$

• Выбор k - интерактивно проверяется практика.

• Выбор Θ - сдвигают или меняют начальные значения для отдельных компонент.

• Обобщение EM-алгоритм предположение - не ~~согласовано~~ согласовано M-step имеет смысл. Там в случаях где есть несколько интервалов, тогда можно ~~получить~~ получить

• Геометрический: Θ может застревать в локальных экстремумах, зависит от первонач. параметров и нелинейно сходится.

$$\text{вместо } \theta_j = \operatorname{argmax} \sum_{i=1}^m g_{ij} \log p(x_i; \theta) \text{ делаем}$$

$$\theta_j = \operatorname{argmax} \sum_{x_i \in V_j} \log p(x_i; \theta)$$

X_j - генерируемые пучки сточаст. моделирования
и обработка $x_i \in X^m$ разделяются на $j^{(i)}$ из дискретного распределения $\{g_{ij}\}$ и обработка. тоо в $j^{(i)}$ входит в выборку $X^{(i)}$.

• С добавл. компонент - син. улучшит. - можно!

(34) Еще

Еще \rightarrow компоненты распределения
но, так как это скрытые. тоо
НЕВЫВОДИМЫ 242%

(34) Выбор признаков.

— признаков может быть слишком много (степени)

— признаки могут быть чрезмерно

- много признаков — это редкое явление

Методы отбора:

- одномерный — оценивает связь признаков с целевым. (корреляцию)
- можно — обходить модели на группах и смотреть где изменяется скор
- LASSO/L1 — регуляризация

Feature selection — GCF, G-выбр. фичи, F-выбр.
Feature extraction — наоборот, G+F (n > k > n)
 $k=161$ $n=1/F$
Хотим выделить независимые признаки

(55) Метод главных компонент

$$z_{ij} = \sum_{k=1}^D w_{ik} x_{ik} = \sum_{k=1}^D x_{ik} w_{ki}^T \Rightarrow z = XW^T$$

Нужно одно 1 реч., нужно $WW^T = I$. Тогда $X = ZW$
и Z будет иметь признаков, поэтому

$$\|X - ZW\|^2 \rightarrow \min_{Z, W}$$

(наши проблемы — сумма квадр. значений)

Алгор. постановка задачи: нам нужно дисперсию наше преобразование, нес аргум.

$$\sum_{j=1}^J w_j^T X^T X w_j \rightarrow \max_w$$

$X^T X$ — ковариационе

• Решение

$$\begin{cases} \sum w_j^T X^T X w_j \rightarrow \max_w \\ w^T w = I \end{cases}$$

Для привед. паралл. сферами без суммы.

$$\begin{cases} w_1^T X^T X w_1 \rightarrow L(w_1, \lambda) = w_1^T X^T X w_1 - \lambda (w_1^T w_1 - 1) \\ w_1^T w_2 = I \end{cases}$$

Дифференцируем лагранжиан: $\frac{\partial L}{\partial w_1} = 0 \Rightarrow X^T X w_1 = \lambda w_1$
 w_2 тоже имеет то же w_1 — ортого. вектор
 $X^T X$ и λ — собств. значение.

онакратко то $w_1^T X^T X w_2 = \lambda$

Данные можно поместить, что нужно брать w_i с макс. значением собств. знач.

Ряде гипотез, сохраняющие: $\sum \lambda_i \leftarrow$ все

$$X = UDV^T - \text{сингл. декомпозиция.}$$

\downarrow матрица W и поган $Z = XW$

Как выглядит λ ? на них используя

• Анализ: маленький, часто ненулев. где близкому, если ненулевые ассими / мультиколлине.

(56) Классификация

Хотим минимиз. выделим. расср. и
вспомог. функции

Цели:

$$F_0 = \frac{\sum [y_i = y_j] f(x_i, x_j)}{\sum [y_i \neq y_j]}$$

\nearrow \rightarrow min

- упрощение дальнейшего обработку данных
- хранение данных — основной предсказатель
- Выделение неодинаковых объектов
- Типизация — классификации

$$\frac{F_0}{F_1} \rightarrow \min$$

Класс. структуры:

- | | | |
|------------|----------------------|--------------------|
| — Гипотезы | — Классиф. с центром | — непересекающиеся |
| — Линии | — с пересечениями | — на один класс |
| | | |

$$\Phi_0 = \sum_{y \in Y} \frac{1}{1k_y} \sum_{i:y_i=y} f^2(x_i, M_y) \rightarrow \min$$

— сумма средних выделим. расср.

$$\Phi_1 = \sum_{y \in Y} f^2(M_y, M) \rightarrow \max$$

$\frac{\Phi_0}{\Phi_1} \rightarrow \min$

Максим. классификация — вероятность классов.

(37) Графовое классификации

Представим выборку как граф. Вершины - узлы, ребра - связи.

- Видение связей наименем пары R - связь $(i, j) / g_{ij} > R$ отсутствует в между R
- Подбором - сортируем матрицу расст.
будет два типа - выбор R
- мин между типами
- Ограничение приведенное
- Плохое управл. типом классов

- Красотин, независимые множества (MST)
 - + наимен. пару верх (i, j) с наим. g_{ij} и соч. пока есть независимые вершины
 - | наимен. вершины в ребрах
 - | удаление $K-1$ самых длинных ребер

Также ограничено применением, $O(\ell^3)$ операции

- FOREL Задано $x_0 \in X$, R
берем все $x_i / g(x_i, x_0) \leq R$ и добавляем в C_1 .
Перемещаем x_0 в центр масс C_1 . Повторяем пока C_1 не добавляется / расчет.
Это приблизительно можно выразить
центром масс $x_i / g(x_i, x_0) \rightarrow \min$
- + Можно определять классы приблиз. пересечениями R
- Чувствителен к x_0

(38) Иерархическое классификации

Также. Стрем. дендрограмм. Для типа
— гибридичное (поглощение меньшее...) и агрегативное (отделение). \leftarrow попарно.

Общем - классиф. Структуры для каждого класса

$$N = UVV$$

$$R(UVV, S) = \frac{d_u R(u, S)}{d_v R(v, S)} + \beta R(u, v) + \gamma |R(u, S) - R(v, S)|$$

Рассмотрим:

- $R^{\text{min}}(W, S) = \min_{w \in W, s \in S} g(w, s) \quad d_u = d_v = \frac{1}{2}, \beta = 0, \gamma = -\frac{1}{2}$
 \uparrow близк. соч.
- $R^{\text{max}}(W, S) = \max_{w \in W, s \in S} g(w, s) \quad -/-, \gamma = \frac{1}{2}$
- $R^{\text{avg}}(W, S) = \frac{1}{|W||S|} \sum_w \sum_s g(w, s) \quad d_u = \frac{|U|}{|W|}, d_v = \frac{|V|}{|S|}, \beta = \gamma = 0$
Среднее
- $R^{\text{max}}(W, S) = g^2 \left(\sum_{w \in W} \frac{w}{|W|}, \sum_{s \in S} \frac{s}{|S|} \right) \quad -/-, \beta = -\alpha_u \alpha_v, \gamma = 0$

R - монотонна, если наше смещение
расст. $\underbrace{\text{меньшее классическое значение ветки}}_{\text{изображающий классами на more}}$

Теорема: Классификация монотонна, если

- 1) $d_u \geq 0, d_v \geq 0$
- 2) $d_u + d_v + \beta \geq 1$
- 3) $\min \{d_u, d_v\} + \gamma \geq 0$

$\Rightarrow R^{\text{max}}$ не монотонна

антиподные б-р не включаются
 $\downarrow \checkmark$ \checkmark \checkmark поглощают
представляющий классами на more

(39) EM-классификации

Гипотеза: X^{ℓ} набл. можно: $\begin{cases} \sum_{y \in Y} w_y p_y(x) = p(x) \\ \sum w_y = 1 \end{cases}$

Гипотеза: $x \in X = \mathbb{R}^n \sim (h \cdot f_n)$, наимен. расст.
или разд. (n -мерный) $p_y(x) \in$ условие $p_y \neq (p_{y1} \cdots p_{ym})$
 $\propto \sum = \text{diag}(\sigma_{y1}^2, \dots, \sigma_{ym}^2)$
 $p_y(x) = (2\pi)^{-\frac{n}{2}} (\sigma_{y1}^2 \cdots \sigma_{ym}^2)^{-\frac{1}{2}} \exp(-\frac{1}{2} \sum_{j=1}^m (x_j - \mu_{yj})^2)$

$$g_y(x, x') = \sum_{j=1}^n b_{yj}^{-2} |f_j(x) - f_j(x')|^2$$

— вычисление евклидово расстояние с весами (b_{yj}^{-2})

Тогда классификация — это разбиение множества. Применение EM-алгоритм.

$$g_{ij} = p(j|x_i) = \frac{w_j p_j(x_i)}{w_1 p_1 + w_2 p_2} \text{ -- вероятн. } x_i$$

Можно мен. EM с нов. начальн. множеством.



• K-means. Дополнение к EM:

1. X принадл. только 1 классу (в ВМ — бинарн.)
2. Несколько находит первые классы

Формирует нач. приближение центров $y \in Y$
 M_y — наименее удаленные дрн от групп
 отвеков

do определите X и бином. центры

$$y_i = \underset{y \in Y}{\arg \min} g(x_i; M_y)$$

бинарн. новое началь. центры

$$M_{yi} = \underset{\text{пока } y_i \text{ меняется}}{\sum_{i=1}^c} \underset{\sum_{i=1}^c}{\underset{y_i=y}{\sum}} f_i(x_i) \quad y \in Y$$

Почему на FOREL с изображ. R.

Чувствительен ко:

— начальным центрам.

— начальным кол-вом классов

Оптимизир. $F_0 \rightarrow \min$, $\Phi_0 \rightarrow \min$

(*) Помощник:

Точки — core (много соседей), border (имеют ^{брз} core — соседи),
 мало), noise — мало соседей. На рисунке ϵ .

DBSCAN: несущие и пасыни.

1. Помогаем точки как связные, граничные или связные
2. Выбрасываем связные
3. Сост. все связн. точки в пачке ϵ — локально связные.
4. Ограничим в пачке
5. Границы — могут пересечь

Погреш. ϵ — погреш. рассм. от KNN и числовых на график.

??