Klasifikatoriaus charakteristikos

Klaidų tipai

Sprendžiant klasifikavimo uždavinį, skiriamos dvejų tipų klaidos:

1. Mokymo klaida tai neteisingai klasifikuotų mokymo imties įrašų dalis.
2. Modelio klaida yra lygi neteisingo naujų įrašų klasifikavimo tikimybei.

Modelio perteklumas

Kai mokymo ir modelio klaidos yra didelės, sakome, kad modelis nepakankamas. Taip gali atsitikti, pavyzdžiui, kai duomenis klasifikuojantis sprendimų medis yra per mažas. Jį didinant mokymo klaida mažėja. Tačiau modelio klaida, labai išplėtus medį, gali pradėti didėti. Kitaip sakant, "per daug gerai" imties duomenis atitinkančio modelio klaida gali būti didesnė nei paprastesnio modelio su didesne mokymo klaida. Tokiu atveju sakome, kad tas sudėtingesnis modelis yra perteklus. Dažniausios perteklumo priežastys:

1) imties nepakankamumas;
2) dalies įrašų iškraipymai;
3) netinkamas modelio konstravimo algoritmo parametrų parinkimas.

Modelio klaidos įverčiai

Todėl reikia konstruoti tokio sudėtingumo modelį, kuris duoda mažiausią modelio klaidą.

Bet kuris klasifikatoriaus konstravimo algoritmas naudoja tik mokymo imties duomenis ir negali tiksliai numatyti kaip sukonstruotas modelis T "elgsis" su naujais įrašais. Kitaip sakant, mes galime tiesiogiai apskaičiuoti mokymo klaidą $e_m = e_m(T)$, bet ne modelio klaidą $e_M = e_M(T)$. Todėl tenka pasitenkinti e_M įverčiais $\hat{e}_M$.

Paprastasis modelio klaidos įvertis

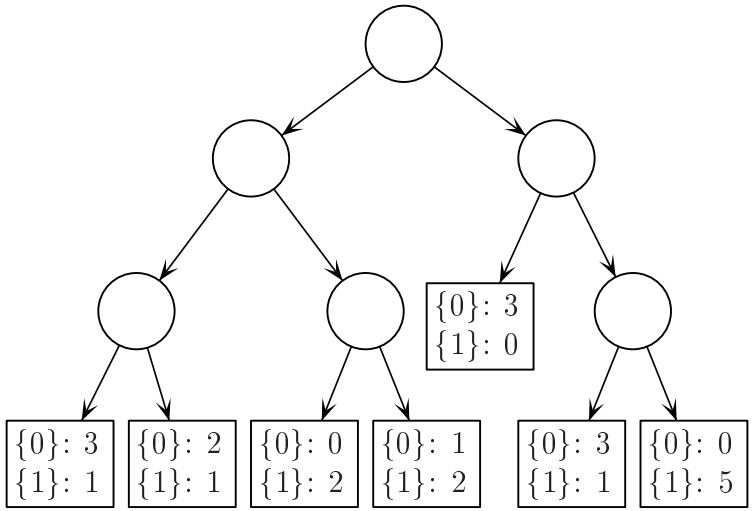
Jei mokymo imtis pakankamai tiksliai atspindi visą populiaciją, tai galime manyti, kad mokymo klaida yra apytiksliai lygi modelio klaidai, t.y.

$$\hat{e}_M = e_m \,.$$

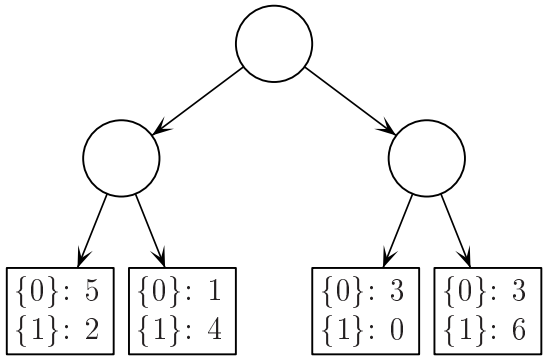
Esant tokiai prielaidai, tiesiog konstruojamas mažiausią mokymo klaidą turintis klasifikatorius.

Paprastasis modelio klaidos įvertis

Pavyzdys. Pagal tą pačią mokymo imtį sukonstruoti du medžiai



Medis T_1



Medis T_2

Turėsime tokius abiejų modelių klaidų įverčius

$$\hat{e}_M(T_1) \quad = \quad e_m(T_1) = \frac{4}{24} \approx 0,1667\,,$$

$$\hat{e}_M(T_2) \quad = \quad e_m(T_2) = \frac{6}{24} = 0,25\,.$$

Pesimistinis modelio klaidos įvertis

Šuo atveju modelio klaidos įvertis gaunamas prie mokymo įverčio pridedant tam tikrą modelio sudėtingumo mokestį. Sprendimų medyje toks mokestis gali būti nustatomas kiekvienam lapui. Tegul medžio T lapai yra $V_1, V_2, \ldots, V_l$. Tada

$$\hat{e}_M(T) = e_m(T) + \frac{1}{N(T)} \sum_{j=1}^{l} \delta(V_j)\,.$$

Čia $N(T)$ - mokymo imties įrašų skaičius, o $\delta(V_j)$ žymi sudėtingumo mokestį lapui V_j.

Pavyzdys. Tegul medžiams T_1 ir T_2 $\delta(V_j) = 0,5$. Tada

$$\begin{aligned} \hat{e}_M(T_1) &= \frac{4}{24} + \frac{1}{24} \cdot 7 \cdot 0,5 = 0,3125\,, \\ \hat{e}_M(T_2) &= \frac{6}{24} + \frac{1}{24} \cdot 4 \cdot 0,5 \approx 0,3333\,. \end{aligned}$$

MDL (Minimum Description Length) įvertis

Tegul duomenys nusakomi atributais $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ ir klasės kintamuoju Y. Sudaryta N įrašų imtis $E = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$. Tarkime Jonas žino visą imtį, o jo geras draugas Petras žino tik atributų reikšmes $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$. Jonas rašo savo draugui laišką, norėdamas pranešti kaip klasifikuojami imties įrašai. Pranešimo ilgis:

1) $L = O(N)$ bitų;
2) jei T klasifikatorius (pavyzdžiui, medis), tai

$$L(E, T) = L(T) + L(E|T)\,.$$

Pagal MDL principą renkamės tą modelį T_{min}, kuriam perduoti reikia trumpiausio pranešimo, t.y.

$$T_{min} = \operatorname*{argmin}_T L(E, T)\,.$$

MDL (Minimum Description Length) įvertis

Pavyzdys. Tarkime, kad medžius T_1 ir T_2 atitinkantys duomenys turi 8 binarinius atributus. Kiekvieną vidinę medžio viršūnę nusakys ją atitinkantis skaidymo atributas. Jis nusakomas $\log_2 8 = 3$ bitais. Šiuo atveju yra tik dvi klasės. Todėl kiekvienas lapas mums kainuos $\log_2 2 = 1$ bitą. Taigi

$$L(T_1) = 6 \cdot 3 + 7 \cdot 1 = 25\,, \quad L(T_2) = 3 \cdot 3 + 4 \cdot 1 = 13$$

Imties E dydis $N = 24$. Todėl

$$\begin{aligned} L(E|T_1) &= e_m(T_1) N \log_2 N = \frac{4}{24} \cdot 24 \cdot \log_2 24 \approx 18,34\,, \\ L(E|T_2) &= e_m(T_2) N \log_2 N = \frac{6}{24} \cdot 24 \cdot \log_2 24 \approx 27,51\,. \end{aligned}$$

Iš čia pagal MDL įvertį išplaukia, kad geresnis yra medis T_2, nes

$$L(E, T_1) \approx 43,34\,, \quad L(E, T_2) \approx 40,51\,.$$

Statistinis įvertis

Tegul klasifikatoriaus T mokymo imtyje yra $N(T)$ įrašų. Tarsime, kad kiekvienas duomenų įrašas, nepriklausomai vienas nuo kito, klaidingai klasifikuojamas su tikimybe $p = e_M(T)$. Rasime šios tikimybės, o tuo pačiu ir modelio klaidos, pasikliautinojo intervalo viršutinį rėžį $e_v(T, Q)$. Tai ir bus statistinis modelio klaidos įvertis

$$\hat{e}_M(T) = e_v(T, Q)\,,$$

čia $Q \in (0,\ 1)$ žymi pasikliovimo lygmenį. Jei $\alpha = \frac{1-Q}{2}$, tai

$$\begin{aligned} e_v(T,Q) &= \left(e_m(T) + \frac{z_\alpha^2}{2N(T)} + z_\alpha \sqrt{\frac{e_m(T)}{N(T)} - \frac{e_m^2(T)}{N(T)} + \frac{z_\alpha^2}{4N^2(T)}} \right) \\ &\times \left(1 + \frac{z_\alpha^2}{N(T)} \right)^{-1}, \end{aligned}$$

čia z_α žymi standartinio normalaus skirstinio $1-\alpha$ lygmens kvantilį.

Kryžminis patikrinimas (Cross-validation)

Mokymui ir kontrolei skirtų imties dalių proporcijos (2/3+1/3).
Sluoksninės (kitaip *stratifikuotos*) imtys.
k - *kartinis kryžminis patikrinimas*(k - fold cross validation): imtis dalijama į k lygių nesikertančių sluoksniuotų dalių. i - toji tampa kontroline imtimi, o likusios $k-1$ dalys sudaro mokymo imtį (paveiksle $k=4,\ i=3$).

Mokymas	Mokymas	Kontrolė	Mokymas

Pagal šią mokymo imtį konstruojamas modelis ir kontrolinėje imtyje apskaičiuojama jo klaida e_i. Galutinis modelio klaidos įvertis $\hat{e}$ yra

$$\hat{e} = \frac{1}{k}\sum_{i=1}^{k} e_i \,.$$

Kai k yra lygus visos imties įrašų skaičiui N, gauname vadinamąjį "vieną išmesk" (leave one out) metodą.

Pakartotinų imčių metodas (bootstrap)

Imtis E sudaryta iš N įrašų. Mokymo imtį E_m sudarome N kartų grąžintinai pasirinkdami imties E įrašą. Kontrolinė imtis $E_k = E \backslash E_m$.

Kiekvienam pradinės imties E įrašui I tikimybė nepatekti į mokymo imtį yra

$$P(I \notin E_m) = \left(1 - \frac{1}{N}\right)^N .$$

Tačiau

$$\lim_{N\to\infty} \left(1 - \frac{1}{N}\right)^N = \frac{1}{\mathrm{e}} \approx 0,368 .$$

Taigi, pakankamai dideliam N kontrolinėje imtyje turėsime apytiksliai $0,368 \cdot N$, o mokymo imtyje atitinkamai $0,632 \cdot N$ pradinės imties E įrašų. Todėl kartais šis metodas dar vadinamas 0,632 pakartotinų imčių (0,632 bootstrap) metodu.

Pakartotinų imčių metodas (bootstrap)

Tarkime, kad aptartoji imčių sudarymo procedūra pakartojama b kartų. Kiekvieną kartą sukonstruojamas atitinkamas klasifikatorius ir jam apskaičiuojamos mokymo ir kontrolinės imties klaidos $e_m(i)$ ir $e_k(i),\ i = 1, 2, \ldots, b$. Tada galutinis modelio klaidos įvertis $\hat{e}_{boot}$ yra

$$\hat{e}_{boot} = \frac{1}{b}\sum_{i=1}^{b}(0,368 \cdot e_m(i) + 0,632 \cdot e_k(i))\,.$$

Pakartotinų imčių metodu randami klaidų įverčiai, kai turima mažai duomenų.

Blogas pavyzdys.
Turime visiškai atsitiktinius įrašus, su tikimybe $0,5$ priklausančius vienai iš dviejų klasių. Konstruojamas pilnai mokymo imtį atsimenantis klasifikatorius. Kitaip sakant, $e_m(i) = 0$. Tuo tarpu $e_k(i) = 0,5$. Todėl $\hat{e}_{boot} = 0,368 \cdot 0 + 0,632 \cdot 0,5 = 0,316\,.$ Akivaizdžiai optimizmo per daug.