ESL 2.6-2.9

Joanna Grunwald

1 Modele statystyczne, nadzorowana nauka, funkcje przybliżające

Naszym celem jest znalezienie funkcji $\hat{f}(x)$, która skutecznie przybliża $f(x)$ i opisuje zależność pomiędzy zmiennymi wejścia i wyjścia. Prowadzi to do funkcji regresji $f(Y|X = x)$. Zbiór najbliższych sąsiadów może być po prostu estymacją z warunkowej wartości oczekiwanej, może mieć to jednak dwie wady. Po pierwsze w przypadku dużego wymiaru przestrzeni wejściowej, najbliżsi sąsiedzi nie muszą znajdować się blisko punktu docelowego i może powodować duże błędy. Drugim przypadkiem jest jeśli wiadomo, że istnieje specjalna struktura, można to wykorzystać do zmniejszenia zarówno skośność i wariancję.

Zakładamy, że nasze dane powstały z modelu statystycznego

$$Y = f(X) + \epsilon,$$

gdzie błąd jest niezależny z X oraz $E(\epsilon) = 0$

1.1 Nadzorowana nauka

Jako pierwszy zaprezentowany zostanie punkt widzenia uczenia maszynowego na dopasowywanie funkcji. Dla uproszczenia zakładamy, że błędy są addytywne oraz model Y jest rozsądnym założeniem. Łatwo opisać to na przykładzie nauczyciela. Obserwujemy badany system, zarówno dane wejściowe, jak i wyjściowe, i tworzy zbiór uczący obserwacji $T = (x_i, y_i), i = 1, \ldots, N$. Obserwowane wartości wejściowe do systemu x_i są również wprowadzane do sztucznego systemu, znanego jako algorytm uczący się. Ten algorytm produkuje wartości wyjściowe $f(x_i)$ w odpowiedzi na dane wejściowe. Może on modyfikować relację $\hat{f}$ w odpowiedzi na różnicę $y_i - \hat{f}(x_i)$. Celem jest, aby po zakończeniu procesu uczenia się, sztuczne i rzeczywiste wyniki były na tyle bliskie, że będą przydatne dla wszystkich zestawów danych wejściowych, które można napotkać w praktyce.

1.2 Funkcje przybliżające

Powyższe podejście jest rozważane w matematyce stosowanej i statystyce z perspektywy aproksymacji funkcji i estymacji. Pary danych (x_i, y_i) są postrzegane jako punkty w $(p + 1)$-wymiarowej przestrzeni euklidesowej. Dla wygody przyjmujemy, że dziedziną jest $\mathbb{R}^p$, a naszym celem jest uzyskanie skutecznego przybliżenia do $f(x)$ dla wszystkich x na pewnym przedziale

$\mathbb{R}^p$.

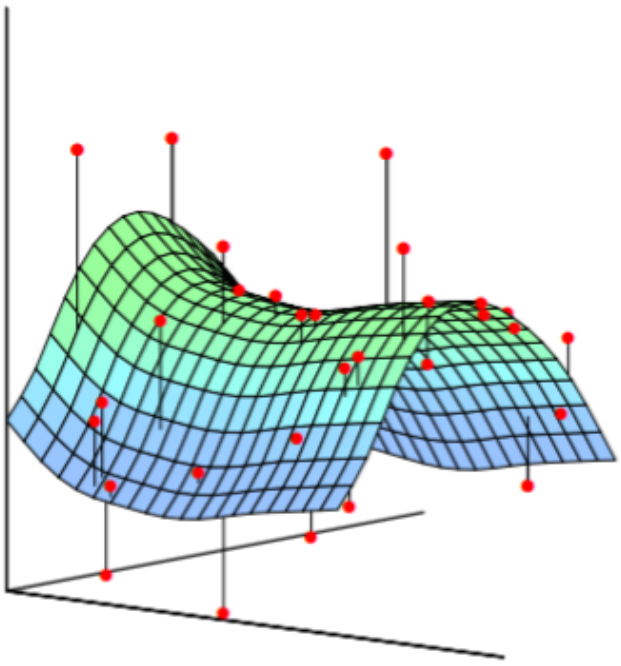
Funkcje przybliżające możemy podzielić na kilka klas, np. rozwinięcia za pomocą liniowej bazy ze zbiorem parametrów θ

$$f_\theta(x) = \sum_{i=1}^{K} h_K(x)\theta_K,$$

gdzie $h_K(x) = \frac{1}{1+exp(-x^T\beta^k)}$ to zbiór funkcji lub transformacji wektora wejściowego x.
Do wyestymowania zbioru θ w f_θ możemy wykorzystać metodę najmniejszych kwadratów, minimalizując sumę reszt

$$RSS(\theta) = \sum_{i=1}^{N}(y_i - f_\theta(x_i))^2,$$

jako funkcję od θ. Jest to rozsądne kryterium w przypadku modelu z addytywnym błędem. Taką sparametryzowaną funkcję traktujemy jako powierzchnię w przestrzeni $\mathbb{R}^{p+1}$, w której obserwujemy zaszumione realizacje. Łatwo to zaobserwować na przykładzie w przestrzeni trójwymiarowej, gdzie oś pionowa opisuje zmienną wyjścia. Szukamy takiego zestawu parametrów θ, żeby dopasowana powierzchnia była jak najbliżej obserwowanych punktów, a bliskość mierzona jest właśnie za pomocą $RSS(\theta)$.



bardziej uogólnionym rozwiązaniem jest szacowanie największej wiarygodności. Na obserwowanej próbce stosujemy następującą metodę

$$L(\theta) = \sum_{i=1}^{N} logPr_\theta(y_i),$$

gdzie $Pr_\theta(y_i)$ to gęstość indeksowana po pewnym parametrze θ. Ta metoda za najbardziej sensowne rozwiązanie uznaje to, dla którego prawdopodobieństwo obserwowanej próbki jest największe.

2 Strukturalne modele regresji

Powyższe metody mogą napotkać problemy w przypadku zwiększenia wymiaru. Dlatego teraz wprowadzimy klasy ustrukturyzowanych podejść. Z każdą klasą jest powiązany jeden lub więcej parametrów, czasami nazywane odpowiednio parametrami wygładzania.

2.1 Kara za nieregularność i metody bayesowskie

Pierwsza klasa funkcji jest kontrolowana przez karę na $RSS(f)$

$$PRSS(f;\lambda) = RSS(f) + \lambda J(f)$$

Odpowiednie $J(f)$ będzie miało duże wartości dla funkcji f, które znacznie się różnią na małych obszarach przestrzeni wejściowej. Funkcje kary J można konstruować dla funkcji w dowolnym wymiarze, a ich specjalne wersje mogą być tworzone w celu narzucenia specjalnej struktury. Np. dodatkowe kary $J(f) = \sum_{j=1}^{p} J(f_j)$ są używane w połączeniu z funkcjami addytywnymi $f(X) = \sum_{j=1}^{p} f_j(X_j)$ do tworzenia modeli addytywnych z gładkimi funkcjami dla współrzędnych. Funkcja kary oraz metody regulowania pokazują, że funkcje, których szukamy, zachowują się, jak funkcje gładkie i zazwyczaj mogą być rzutowane w ramach bayesowskich.

Kara J pochodzi z rozkładu apriori, a $PRSS(f;\lambda)$ aposteriori, zatem zminimalizowanie $PRSS(f;\lambda)$ sprowadza się do znalezienia trybu post.

2.2 Metody jądra i regresja lokalna

Te metody szacują funkcje regresji lub warunkowej wartości oczekiwanej poprzez określenie charakteru sąsiedztwa lokalnego, a także za pomocą klas funkcji dopasowanych lokalnie. Natomiast lokalne sąsiedztwo jest określone przez funkcję jądra np. dla jądra Gaussa ma funkcję wagi opartą na gęstości

$$K_\lambda(x_0, x) = \frac{1}{\lambda} exp\left(\frac{||x - x_0||^2}{2\lambda}\right).$$

Parametr λ odpowiada wariancji i kontroluje szerokość otoczenia. Najprostszą formą oszacowania jądra jest średnia ważona Nadaraya-Watsona opisana wzorem

$$\hat{f}(x_0) = \frac{\sum_{i=1}^{N} K_\lambda(x_0, x_i) y_i}{\sum_{i=1}^{N} K_\lambda(x_0, x_i)}$$

Zdefiniujemy $f_{\hat{\theta}}(x_0)$ jako funkcję przybliżającą $f(x_0)$, gdzie $\hat{\theta}$ minimalizuje

$$RSS(f_\theta, x_0) = \sum_{i=1}^{N} K_\lambda(x_0, x_i)(y_i - f_\theta(x_i))^2,$$

a f_θ jest sparametryzowaną funkcją, np. wielomianem niskiego rzędu.

Metodę najbliższego sąsiada można traktować jako metodę jądra z rozszerzeniem metryki, przez co jest bardziej zależna od danych. W rzeczywistości, metryka k-najbliższych sąsiadów to

$$K_k(x, x_0) = I(||x - x_0|| \leq ||x_{(k)} - x_0||),$$

gdzie $x_{(k)}$, to k-ty element, sortując po odległości od x_0.
W wyższych wymiarach te metody muszą być zmodyfikowane.

2.3 Funkcje bazowe

Ta klasa metod obejmuje znane rozwinięcia liniowe i wielomianowe, ale co ważniejsze, szeroką gamę hatdziej elastycznych modeli. Model dla funkcji f jest liniowym wyrażeniem na funkcjach bazowych

$$f_\theta(x) = \sum_{m=1}^{M} \theta_m h_m(x),$$

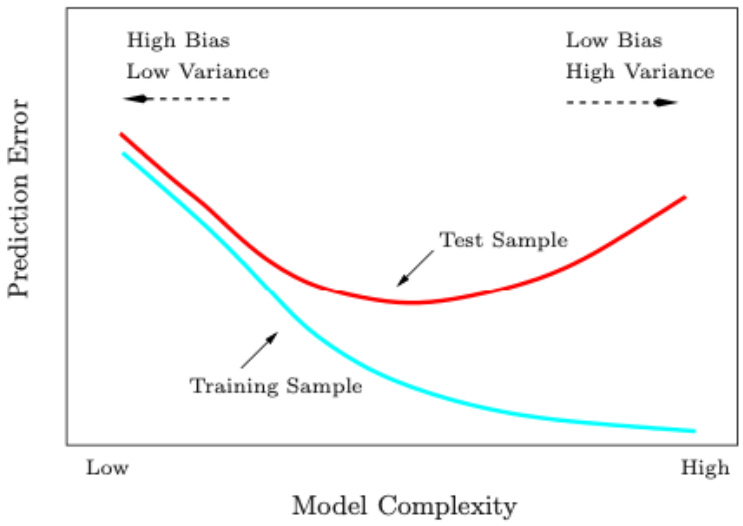
gdzie h_m jest funkcją wejścia dla x, a parametry θ są liniowe.
Dla jednowymiarowego wektora x, wielomianowe splajny stopnia K mogą być prezentowane jako sekwencja M splajnów bazowych z $M-K$ węzłami. W wyniku pomiędzy węzłami dostajemy funkcje, które są wielomianami stopnia K i łączą się ciągle.
Jako przykład możemy rozważać funkcję liniowe na odcinkach między węzłami. Intuicyjną bazą jest wtedy $b_1(x) = 1$, $b_2(x) = x$, $b_{m+2}(x - t_m)_+$, $m = 1, \ldots, M-2$, gdzie t_m to m-ty węzeł, a z_+ oznacza część dodatnią. Wtedy parametr θ może być całkowitym stopniem wielomianu lub w przypadku splajnów liczbą węzłów.

3 Wybór modelu oraz kompromis obciążenia i wariancji

Wszystkie modele wyżej opisanie miało parametr wygładzania lub złożoności, m.in. mnożnik wymiaru kary, szerokość jądra, czy ilość funkcji bazowych.
Dopasowanie regresji k-najbliższego sąsiada $f_k(x_0)$ pokazuje siły celownicze, które wpływają na zdolność przewidywania takich przybliżeń. Załóżmy, że dane pochodzą z modelu $Y = f(X) + \epsilon$, gdzie $E(\epsilon) = 0$ i $Var(\epsilon) = \sigma^2$. Dla uproszczenia zakładamy, że wartości x_i w próbce są z góry ustalone. Wtedy oczekiwany błąd prognozowania w punkcie x_0, znany również jako błąd testowy wyrażamy następująco

$$\begin{aligned} EPE_k(x_0) &= E[(Y - \hat{f}_k(x_0))^2 | X = x_0] \\ &= \sigma^2 + [Bias^2(\hat{f}_k(x_0)) + Var(\hat{f}_k(x_0))] \\ &= \sigma^2 + [f(x_0) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})]^2 + \frac{\sigma^2}{2} \end{aligned}$$

Pierwszy składnik σ^2 jest nieredukowalny. Mamy natomiast wpływ na drugi i trzeci składnik, czyli obciążenie i wariancję. W przypadku obciążenia wartość oczekiwana uśrednia losowość danych uczących się. W większości przypadków ta wartość będzie przyrastać razem z k. Dla małego k kilku najbliższych sąsiadów będzie miało wartości $f(x_{(l)})$ bliskie $f(x_0)$, więc ich średnia powinna być bliska $f(x_0)$. Gdy k rośnie, sąsiedzi są coraz dalej, a wtedy wszystko może się zdarzyć.



Wariancja maleje jako odwrotność k. Zatem, gdy k się zmienia, istnieje kompromis między obciążeniem, a wariancją.