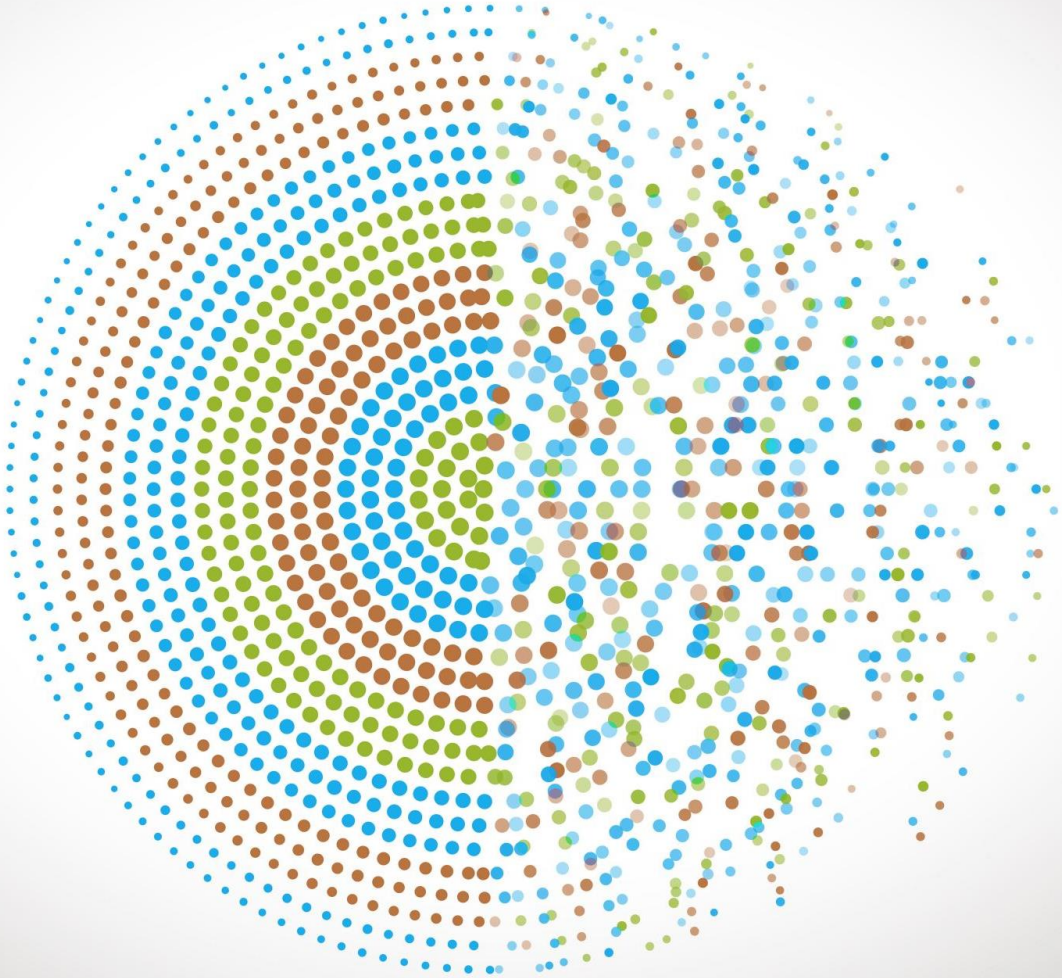
Problemy, w których liczba cech jest dużo większa niż liczba obserwacji

Aleksandra Siepiela

Czerwiec 2022



Problemy, w których p>>N

W tym rozdziale zajmiemy się problemami predykcji, w których liczba cech p jest dużo większa niż liczba obserwacji N (p >> N). Takie problemy często pojawiają się szczególnie w genomice i innych obszarach biologii obliczeniowej.

Głównymi problemami, które pojawiają się w zagadnieniach tego typu są wysoka wariancja i overfitting. W związku z tym, w sytuacjach, w których p >> N najczęściej wybiera się proste podejścia regularyzowane.

Przykład cz.1

Na początku rozważmy obrazek 18.1. Rysunek ten podsumowuje małe badanie symulacyjne, które pokazuje, że w przypadku problemów, w których $p >> N$, zasada "mniejsze dopasowanie jest lepsze" (z ang. "less fitting is better") ma zastosowanie.

W trakcie badania dla każdej z $N = 100$ próbek wygenerowano p cech X ze standardowego rozkładu Gaussowskiego. Pomiędzy każdą parą zmiennych występuje korelacja na poziomie 20%. Wyniki Y zostały wygnerowane zgodnie z modelem liniowym:

$$Y = \sum_{j=1}^{p} X_j \beta_j + \sigma\varepsilon \tag{18.1}$$

gdzie ϵ został wygenerowany ze standardowego rozkładu Gaussowskiego.

Przykład cz.2

Dla każdego zbioru danych, zbiór współczynników β_j również został wygenerowany z rozkładu standardowego Gaussowskiego. Rozważymy trzy przypadki $p = 20, 100, 1000$. Odchylenie standardowe, w każdym przypadku zostało dobrane tak, aby stosunek sygnału do szumu $\frac{Var(E(Y|X))}{\sigma^2}$ był równy 2. Liczba istotnych współczynników regresji to odpowiednio 9, 33, 331 (uśredniając dla 100 symulacji).

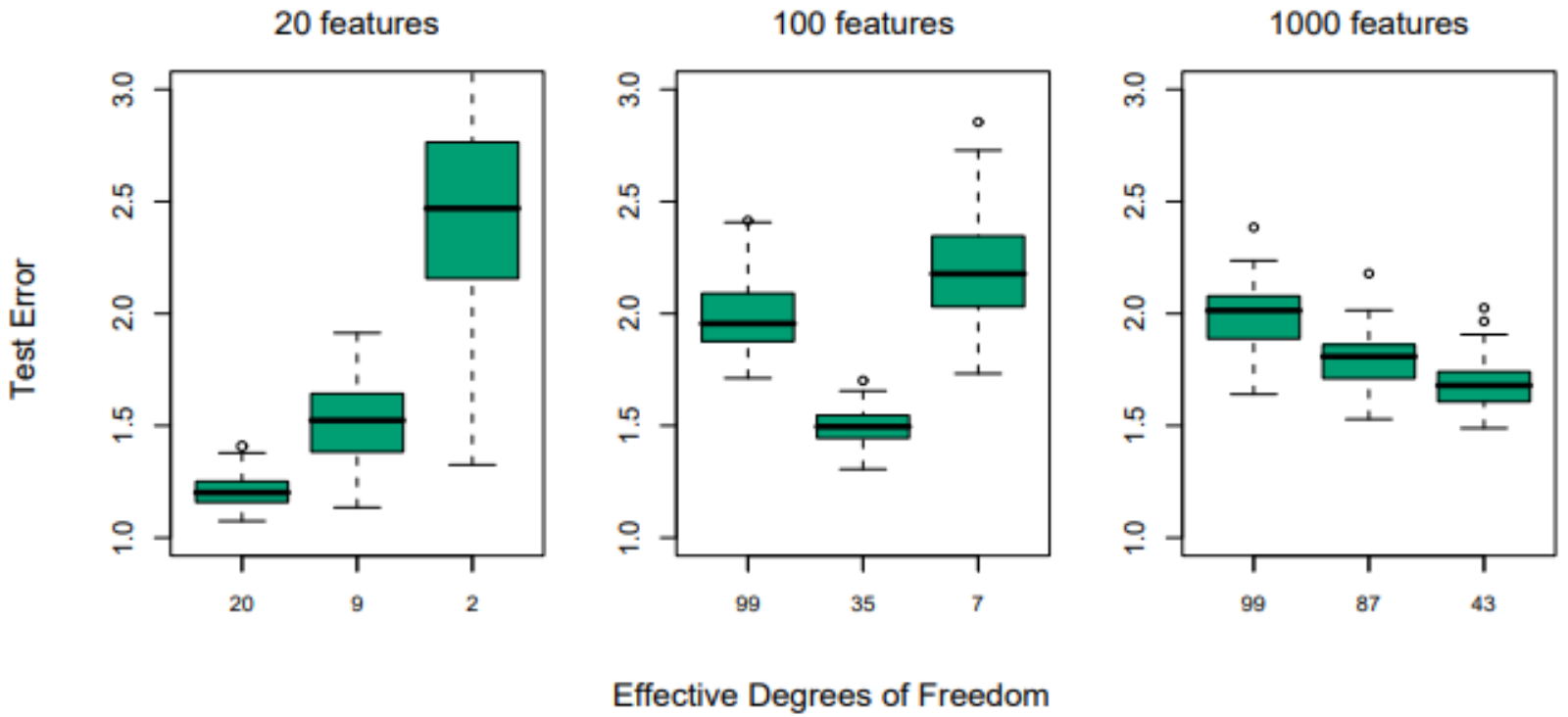


FIGURE 18.1. *Test-error results for simulation experiments. Shown are box-plots of the relative test errors over* 100 *simulations, for three different values of* p*, the number of features. The relative error is the test error divided by the Bayes error,* σ^2*. From left to right, results are shown for ridge regression with three different values of the regularization parameter* λ*:* 0.001, 100 *and* 1000*. The (average) effective degrees of freedom in the fit is indicated below each plot.*

Podsumowanie wyników

Analizując otrzymane wyniki możemy zauważyć, że w sytuacji, gdy $p = 20$, najmniejszy względny błąd testowy obserwujemy dla regresji grzbietowej z $\lambda = 0.001$ (20 df), gdy $p = 100$ najlepszy wynik dostajemy dla regresji grzbietowej z $\lambda = 100$ (35 df), natomiast, gdy $p = 1000$ dla regresji grzbietowej z $\lambda = 1000$ (43 df) .

Skąd taki wynik?

Regresja grzbietowa z $\lambda = 0.001$ dobrze radzi sobie z korelacją pomędzy cechami, gdy $p < N$, ale nie w przypadku, gdy $p >> N$. W późniejszych przypadkach nie mamy dostatecznie dużej ilości informacji w relatywnie małej liczbie próbek, by efektywnie wyestymować wielowymiarową macierz kowariancji. W tym przypadku, większa regularyzacja prowadzi do lepszej wydajności przewidywania. Nic więc dziwnego, że analiza danych wielowymiarowych wymaga albo modyfikacji procedur zaprojektowanych dla scenariusza $N > p$, albo wprowadzenia całkowicie nowych procedur.

Diagonal Linear Discriminant Analysis

Diagonal Linear Discriminant Analysis

Metoda, którą teraz przedstawię jest zbliżona do metod dyskutowanych w rozdziale 4.3.1 książki *The Elements of Statistical Learning*. W tym wypadku jednak wprowadzono pewną modyfikację, która pozwala osiągnąć selekcję zmiennych.

Najprostsza forma regularyzacji zakłada, że cechy są niezależne w ramach każdej klasy, czyli wewnątrzklasowa macierz kowariancji jest diagonalna.

W przypadku, gdy $p >> N$:

- cechy w obrębie klasy rzadko są niezależne;
- nie mamy dostatecznie dużo danych by wyestymować zależności pomiędzy nimi.

Diagonal Linear Discriminant Analysis

Założenie niezależności zmiennych w dużym stopniu redukuje liczbę parametrów w modelu i skutkuje otrzymaniem efektywnego i łatwo interpretowalnego klasyfikatora.

Dlatego bierzemy pod uwagę regułę LDA diagonalnej kowariancji do klasyfikowania klas. Score dyskryminacyjny dla klasy k możemy wyrazić następująco:

$$\delta_k(x^*) = -\sum_{j=1}^{p} \frac{(x_j^* - \bar{x}_{kj})^2}{s_j^2} + 2\log \pi_k. \qquad (18.2)$$

W sytuacji, gdy rozważalibyśmy dane dotyczące genów, elementy powyższego wzoru interpretowalibyśmy następująco:

$x^* = (x_1^*, ..., x_p^*)^T$ – wektor wartości ekspresji genów dla obserwacji testowej;
s_j – odchylenie standardowe j-tego genu w obrębie klasy;
$\overline{x}_{kj} = \sum_{i \in C_k} x_{ij}/N_k$ – średnia N_k wartości dla genu j w klasie k;
C_k – zbiór indeksów klasy k.

Diagonal Linear Discriminant Analysis

Wyrażenie $\bar{x}_k = (\bar{x}_{k1}, ..., \bar{x}_{kp})^T$ nazywane jest **centroidem** klasy k.

Zauważymy również, że pierwsza (ujemna) część wzoru 18.2 jest po prostu znormalizowaną kwadratową odległością x^* do k-tego centroidu, natomiast druga część wzoru jest poprawką na podstawie klasowego prawdopodobieństwa *prior* π_k, gdzie $\sum_{k=1}^{K} \pi_k = 1$.

Znając już powyższe fakty możemy przedstawić następującą regułę klasyfikacji:

$$C(x^*) = \ell \text{ if } \delta_\ell(x^*) = \max_k \delta_k(x^*). \qquad (18.3)$$

Widzimy, że klasyfikator diagonal LGD odpowiada klasyfikatorowi najbliższego centroidu po odpowiedniej standaryzacji. Klasyfikator diagonal LDA jest również specjalnym przypadkiem naiwnego klasyfikatora Bayesa.

Diagonal Linear Discriminant Analysis

Klasyfikator diagonal LDA często sprawdza się w sytuacji, gdy rozważamy przestrzenie wielowymiarowe. Jednakże, jedną z jego wad jest to, że klasyfikator ten wykorzystuje wszystkie cechy i w związku z tym nie jest wygodny do interpretacji.

Linear Classifiers with Quadratic Regularization

Regularized Discriminant Analysis, gdy p<N

W roku 1989 Friedman zaproponował pewien kompromis pomiędzy metodą LDA i QDA - RDA. Metoda ta pozwala na wprowadzenie pewnej regularyzacji macierzy kowariancji. Regularyzowane macierze kowariancji mają wówczas następującą postać:

$$\hat{\mathbf{\Sigma}}_k(\alpha) = \alpha\hat{\mathbf{\Sigma}}_k + (1-\alpha)\hat{\mathbf{\Sigma}}, \qquad (4.13)$$

gdzie $\hat{\Sigma}$ jest połączoną macierzą kowariancji tak jak w LDA, $\alpha \in [0,1]$ jest stałą. W prkatyce α może zostać wybrana np. przy pomocy walidacji krzyżowej.

Regularized Discriminant Analysis, gdy p>>N

Metoda RDA została opisana w sekcji 4.3.1 w książce *The Elements of Statistical Learning*.

Liniowa analiza dyskryminacyjna wymaga odwrócenia macierzy kowariancji $p \times p$. Gdy $p >> N$ taka macierz może być ogromna, ma rząd co najwyżej $N < p$ i stąd też jest osobliwa. Metoda RDA pokonuje problemy z osobliwością poprzez regularyzację estymatora $\hat{\Sigma}$. Poniżej przedstawimy wersję RDA, która ściąga $\hat{\Sigma}$ w stronę diagonali:

$$\hat{\mathbf{\Sigma}}(\gamma) = \gamma\hat{\mathbf{\Sigma}} + (1-\gamma)\text{diag}(\hat{\mathbf{\Sigma}}), \text{ with } \gamma \in [0,1]. \qquad (18.9)$$

Regularized Discriminant Analysis, gdy p>>N

Forma ściągania w przypadku (18.9) przypomina regresję grzbietową, która ściąga macierz kowariancji cech w stronę diagonalnej macierzy. Wartość γ wybierana jest przy pomocy walidacji krzyżowej, natomiast metody, które służą do odwrócenia tej macierzy zostały omówione w rozdziale 18.3.5 w książce The Elements of Statistical Learning.

Logistic Regression with Quadratic Regularization

Gdy rozważamy problem $p >> N$, regresję logistyczną możemy zmodyfikować w podobny sposób. Rozważmy K klas i wieloklasowy model regresji logistycznej:

$$\Pr(G = k|X = x) = \frac{\exp(\beta_{k0} + x^T \beta_k)}{\sum_{\ell=1}^{K} \exp(\beta_{\ell 0} + x^T \beta_\ell)}. \qquad (18.10)$$

W przedstawionym powyżej modelu rozważamy K wektorów współczynników $\beta_1, \beta_2, ..., \beta_K$.

Logistic Regression with Quadratic Regularization

W przypadku regresji logistycznej, gdy $p >> N$, regularyzacji dopasowania dokonujemy poprzez nałożenie odpowiedniej kary i maksymalizację otrzymanego wyrażenia:

$$\max_{\{\beta_{0k},\beta_k\}_1^K} \left[\sum_{i=1}^{N} \log \Pr(g_i|x_i) - \frac{\lambda}{2} \sum_{k=1}^{K} ||\beta_k||_2^2 \right]. \qquad (18.11)$$

Regularyzacja automatycznie rozwiązuje nadmiarowowść w parametryzacji i wymusza $\sum_{k=1}^{K} \hat{\beta}_{kj} = 0$ dla $j = 1, ..., p$. Warto również zwrócić uwagę, że intercepty β_{k0} nie są regularyzowane (dlatego należy je ustawić na zero).

Rozważany problem optymalizacyjny jest wypukły i może zostać rozwiązany przy pomocy algorytmu Newtona lub innych technik numerycznych. Szczegóły przedstawione zostały w artykule Zhu i Hastiego z roku 2004.

The Support Vector Classifier

Metoda The support vector classifier została opisana dla przypadku dwóch klas w rozdziale 12.2 książki *The Elements of Statistical Learning*.

Gdy rozważamy problem, w którym $p >> N$, metoda ta jest szczególnie przydatna, ponieważ klasy mogą zostać dokładnie rozdzielone przy pomocy hiperpłaszczyzny, chyba że w różnych klasach znajdują się identyczne wektory cech.

Co ciekawe, w przypadku problemów, w których $p >> N$, nieregularyzowany support vector classifier często działa tak samo dobrze jak jego regularyzowana wersja. Jest to związane z tym, że w przypadku tego klasyfikatora nadmierne dopasowanie do danych często nie stanowi problemu.

The Support Vector Classifier

W roku 2007, Tibshirani i Hastie zaproponowali model *margin tree classifier*, w którym klasyfikatory support-vector są wykorzystywane w drzewie binarnym. Klasy są zorganizowane hierarchicznie, co może być przydane na przykład przy klasyfikowaniu pacjentów z różnym typem raka.

Dziękuję za uwagę