Skorelowanie cech
Analiza głównych składowych

Korelacja cech

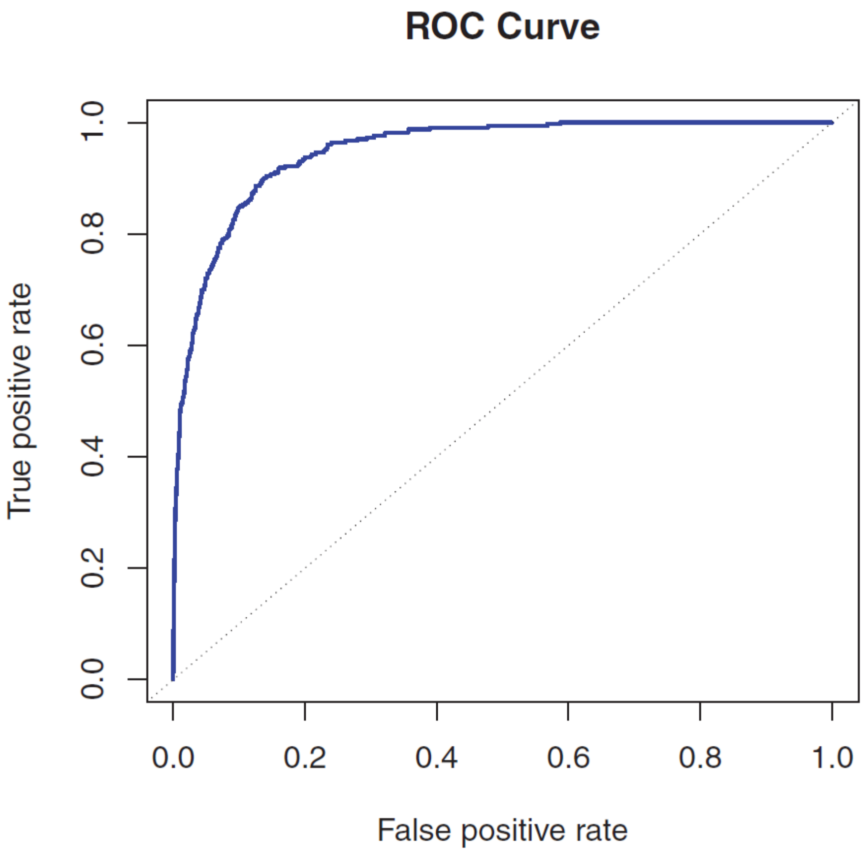
Współczynnik korelacji liniowej Pearsona

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Współczynnik korelacji rang Spearmana

$$\rho_{X,Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}$$

Receiver Operating Characteristic Curve



		Forecast		
		0	1	
Actuals	0	TN	FP	N
	1	FN	TP	P
		N*	P*	

$$False\ positive\ rate = \frac{FP}{N}$$

$$True\ positive\ rate = \frac{TP}{P}$$

Variance Inflation Factor

- k zmiennych objaśniających
- Estymujemy model dla zmiennej i:

$$X_i = \alpha_1 X_1 + \cdots + \alpha_{i-1} X_{i-1} + \alpha_{i+1} X_{i+1} + \cdots + \alpha_k X_k + \alpha_0 + \varepsilon$$

- Obliczamy dla każdej zmiennej

$$VIF_i = \frac{1}{1 - R_i^2}$$

- Jeżeli VIF_i jest większy od pewnego progu to uznajemy, że zmienna jest liniowo zależna od pozostałych zmiennych

Principal Component Analysis

Twierdzenie

Jeżeli A oraz B są macierzami symetrycznymi (a macierz B jest odwracalna) to maksimum $x^T A x$ pod warunkiem $x^T B x$ = 1 jest równe największej wartości własnej macierzy $B^{-1}A$:

$$\max\ x^T A\, x = \lambda_1 \geq \cdots \geq \lambda_p = \min\ x^T A x.$$

Principal Component Analysis

- Tworzymy kombinacje liniowe cech $\delta^T X$, gdzie $\|\delta\| = 1$
- Celem jest maksymalizacja wariancji $\delta^T X$
- Maksymalizacja $V(\delta^T X) = \delta^T \Sigma \delta$, gdzie $V(X) = \Sigma$, prowadzi do wyboru $\delta = \gamma_1$, pierwszego wektora własnego macierzy wariancji Σ odpowiadającego największej wartości własnej λ_1 (→poprzedni slajd)
- Powyższe jest uogólniane na wyższe wymiary – istnieje p wektorów własnych i p głównych składowych $Y_1, \dots, Y_p$.
- Kombinację $Y_i = X\gamma_i$ nazywamy i-tą główną składową (PC_i)
- $Y = X\Gamma, \qquad \Gamma = (\gamma_1, \dots, \gamma_p)$
- $V(X) = \Sigma = \Gamma^T \Lambda \Gamma = \Gamma \Lambda \Gamma^T$

Principal Component Analysis

- Właściwości głównych składowych:
  - $E(Y_j) = 0$
  - $V(Y_j) = \lambda_j$
  - brak skorelowania
  - $\sum_{j=1}^{p} V(Y_j) = tr(\Sigma) = |\Sigma|$
  - $\lambda_1 \geq \cdots \geq \lambda_p \Rightarrow V(Y_1) \geq \cdots \geq V(Y_p)$
- Skala cech powinna być zbliżona
- Objaśnienie wariancji całkowitej przez q pierwszych PC:

$$\psi_q = \sum_{j=1}^{q} \lambda_j \Big/ \sum_{j=1}^{p} \lambda_j$$

Principal Component Analysis

- $Cov(X,Y) = \Gamma\Lambda$
- $\rho_{X_i Y_j} = \gamma_{ij}\left(\frac{\lambda_j}{\sigma_{X_i X_i}}\right)^{1/2}$
- Wybór liczby PC – próg, wartości własne, wykres ψ
- Zastosowania – redukcja wymiaru danych, rangowanie, wizualizacja do trzech wymiarów, sposób radzenia sobie z „przekleństwem wielowymiarowości", punkt wyjścia dla innych metod