Zestaw 3 - Zadanie 2

Metody probabilistyczne w uczeniu maszynowym

Łukasz Trzos

Treść zadania

Niech kolumny x_j, gdzie $j = 1, ..., k$, macierzy planowania X o wymiarach $m \times k$ będą wektorami ortonormalnymi (czyli $m \geqslant k$). Niech $\hat{\theta}$ będzie rozwiązaniem problemu regresji z funkcją kwadratową bez regularyzacji. Wykaż, że rozwiązanie problemu regresji lasso z parametrem regularyzacji λ jest postaci

$$\hat{\theta^l_\lambda} = \mathrm{sgn}(\hat{\theta})\mathrm{max}\{|\hat{\theta}| - \frac{\lambda}{2}, 0\}$$

Rozwiązanie

Z zadania pierwszego wnioskujemy, że rozwiązanie problemu regresji grzbietowej jest postaci:

$$\hat{\theta^r} = (X^TX + \lambda\mathbf{I})^{-1}X^Ty$$

W naszym przypadku parametr regularyzacji wynosi 0. Ponadto, macierz X jest ortogonalna, zatem $X^TX = \mathbf{I}$. Po podstawieniu otrzymujemy:

$$\hat{\theta^l_\lambda} = X^Ty$$

Rozwiązaniem problemu regresji lasso jest argument minimalizujący sumę:

$$||X\theta - y||_2^2 + \lambda||\theta||_1$$

Obliczamy:

$$||X\theta - y||_2^2 + \lambda||\theta||_1 = (X\theta - y)^T(X\theta - y) + \lambda||\theta||_1 = (\theta^TX^TX\theta - 2\theta^TX^Ty + y^Ty) + \lambda||\theta||_1 =$$

$$= (\theta^T\theta - 2\theta\hat{\theta} + \hat{\theta}^TX^TX\hat{\theta}) + \lambda||\theta||_1 = (\theta^T\theta - 2\theta\hat{\theta} + \hat{\theta}\hat{\theta}) + \lambda||\theta||_1 = ||\theta - \hat{\theta}||_2^2 + \lambda||\theta||_1$$

Obliczamy jak pojedynczy składnik wektora θ kontrybuuje do tej sumy:

$$\frac{d}{d\theta_i}||\theta - \hat{\theta}||_2^2 + \lambda||\theta||_1 = 2(\theta_i - \hat{\theta}_i) + \lambda\mathrm{sgn}(\theta_i)$$

Jeśli $\theta_i > 0$ to wartość pochodnej wynosi $2(\theta_i - \hat{\theta}_i) + \lambda$

Optymalnym współczynnikiem jest więc $\theta_i = \hat{\theta}_i - \frac{\lambda}{2}$. Nie zawsze mieści się on jednak w dziedzinie. Jeśli pochodna zerowałaby się dla ujemnego współczynnika, a wiemy, że suma rośnie do nieskończoności wraz ze zwiększaniem parametru, to najmniejszą sumę daje nam brzeg przedziału najbliżej miejsca zerowego pochodnej, czyli 0. Zatem $\theta_i = \max\{\hat{\theta}_i - \frac{\lambda}{2}, 0\}$.

Analogiczne obliczenia możemy poprowadzić dla $\theta_i < 0$ i otrzymamy symetryczny wynik. Zatem jeśli przez $\mathrm{sgn}(\hat{\theta})$ rozumiemy wektor znaków jego składników, to rozwiązanie problemu danego w zadaniu ma postać

$$\mathrm{sgn}(\hat{\theta})\mathrm{max}\{|\hat{\theta}| - \frac{\lambda}{2}, 0\}$$

$\square$