Algorytmy Statystyki Praktycznej

Błażej Miasojedow

MIMUW

28 lutego 2019

Podstawowe informacje.

- Błażej Miasojedow, email: bmiasojedow@mimuw.edu.pl
- Konsultacje po wykładach.
- www.mimuw.edu.pl/∼bmia/asp

Zasady zaliczenia przedmiotu.

- Laboratorium : 20 pkt.
- Egzamin : 20 pkt.
- Ocena końcowa : Laboratorium + Egzamin

Egzamin odbędzie się na ostatnim wykładzie.

O czym będzie wykład?

Większość metod statystycznych w praktyce wymag dodatkowych metod obliczeniowych, pozwalających na efektywne wyznaczanie estymatorów lub innych wielkości w realnych problemach.

1. W staystyce częstościowej, najpopularnijeszym o dobrych własnościach estymatorem jest ENW postaci:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathrm{L}(\theta).$$

2. W statystyce bayesowskiej estymatory są postaci:

$$\hat{\theta} = \mathbb{E}(\theta|\mathrm{X}),$$

gdzie typowo nie można wyznaczyć analitycznie rozkładu $\theta|\mathrm{X}$.

O czym będzie wykład?

Zapoznamy sie z metodami statystyki obliczeniowej ich własnościami i ograniczeniami. Wykład będzie sie składał z 3 podstawowych części:

1. Optymalizacja (wypukła).
2. Metody MCMC.
3. Wnioskowanie dla ukrytych modeli Markowa.

Optymalizacja w statystyce

1. Estymatory najwiekszej wiarygodności
2. Regresja liniowa.
3. Regresja logistyczna (Uogólnione modele liniowe)
4. Maszyny wektorów podpierających
5. ...

Dlaczego warto zrozumieć algorytmy?

Dla większości problemów statystyki istnieją efektywne gotwe implementacje w popularnych językach. Dlaczego warto się uczyć jakie w nich są ukryte metody?

- Nie istnieją globalnie najlepsze metody, dla konkretnego problemu jedne metody są efektywne, a dla innego mogą zawieść.
- Często wiedza o metodach rozwiązywania problemu optymalizacyjnego pozwala na lepsze zrozumienie własności estymatorów (np. wiele wyników dla LASSO zostało uzyskanych dzięki analizie warunków KKT)
- Zrozumienie metod obliczeniowych pozwala na unikanie budowania zbyt skomplikowanych modeli.

Regresja liniowa

Jednym z najpopularniejszych modeli statystycznych jest model liniowy, czyli

$$\mathrm{Y} = \mathrm{X}\beta + \varepsilon \ ,$$

gdzie

- $\mathrm{Y} \in \mathbb{R}^{\mathrm{n}}$ – wektor cech objaśnianych;
- $\mathrm{X} \in \mathbb{R}^{\mathrm{n}\times\mathrm{p}}$ – macierz planu;
- β - wektor nieznanych parametrów;
- ε - niezależny szum o wariancji σ^2.

Metoda najmniejszych kwadratów

Chcemy zminimalizować odległość

$$\mathrm{RSS}(\beta) = \|\mathrm{Y} - \mathrm{X}\beta\|^2$$

Po przyrównaniu gradientu do zera otrzymujemy

$$\beta = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$$

Czy to jest pełna odpowiedź jak obliczyć β w praktyce?

Metoda najmniejszych kwadratów

Chcemy zminimalizować odległość

$$\mathrm{RSS}(\beta) = \|\mathrm{Y} - \mathrm{X}\beta\|^2$$

Po przyrównaniu gradientu do zera otrzymujemy

$$\beta = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$$

Czy to jest pełna odpowiedź jak obliczyć β w praktyce?

Złożoność obliczeniowa

Oblicza się w (przybliżeniu) potrzebną do rozwiązania liczbę podstawowych operacji arytmetycznych:

Operacje wektorowe: $a, b \in \mathbb{R}^n$

- Dodanie dwóch wektorów $a + b$ ma koszt n.
- Iloczyn skalarny $a^T b$ ma koszt $2n$.
- mnożenie przez skalar $c \cdot a$ ma koszt n.

Operacje macierzowe: $A \in \mathbb{R}^{n \times m}, B \in \mathbb{R}^{n \times p}, c \in \mathbb{R}^p$

- Bc ma koszt $2np$
- AB ma koszt $2mnp$
- ABc poprawnie liczone ma koszt $2np + 2mn$, niepoprawnie $2mnp + 2np$

Rozwiązywanie równań liniowych

Chcemy numerycznie rozwiązać równanie $\mathrm{Ax} = \mathrm{b}$, czyli $\mathrm{x} = \mathrm{A}^{-1}\mathrm{b}$, gdzie $\mathrm{A} \in \mathbb{R}^{\mathrm{n}\times\mathrm{n}}$.

- W ogólności problem ten jest rozwiązywany w koszcie n^3.
- Jeżeli A diagonalna to w koszcie n.
- Jeżeli A (górno/dolno)trójkątna w koszcie n^2

Rozkład Choleskiego

Jeżeli macierz A jest symetryczna i dodatnio określona to istnieje jej rozkład

$$A = LL^T ,$$

gdzie L dolnotrójkątna. Koszt obliczenia L jest w przybliżeniu $n^3/3$

Rozwiązywanie równania liniowego za pomocą rozkładu Choleskiego

- $LL^Tx = b$, podstawiamy $y = L^Tx$
- rozwiązujemy równanie $Ly = b$.
- rozwiązujemy równanie $L^Tx = y$

Łączny koszt to $n^3/3 + 2n^2$.

Rozkład Choleskiego

Jeżeli macierz A jest symetryczna i dodatnio określona to istnieje jej rozkład

$$A = LL^T ,$$

gdzie L dolnotrójkątna. Koszt obliczenia L jest w przybliżeniu $n^3/3$

Rozwiązywanie równania liniowego za pomocą rozkładu Choleskiego

- $LL^Tx = b$, podstawiamy $y = L^Tx$
- rozwiązujemy równanie $Ly = b$.
- rozwiązujemy równanie $L^Tx = y$

Łączny koszt to $n^3/3 + 2n^2$.

MNK a rozkład Choleskiego

$$\min_{\beta} \|y - X\beta\|^2$$

Koszt za pomocą rozkładu Choleskiego

- $X^T y$ koszt $2np$
- $X^T X$ koszt np^2.
- Rozkład Choleskiego $X^T X$ koszt $p^3/3$
- Rozwiązanie $X^T X\beta = X^T y$ koszt $2p^2$

Łącznie $np^2 + p^3/3$, jeśli $n >> p$ to koszt np^2.

Rozkład QR

Niech $A \in \mathbb{R}^{n\times p}$ gdzie $n \geqslant p$, wówczas istnieje rozkład

$$A = QR\ ,$$

gdzie $Q \in \mathbb{R}^{n\times p}$ ortogonalna czyli $Q^T Q = I$ oraz R górnotrójkątna. Koszt obliczenia rozkładu QR w przybliżeniu wynosi $2(n - p/3)p^2$.
Macierz Q jest bazą przestrzeni liniowej rozpiętej przez kolumny macierzy A. Ponadto dla każdego x zachodzi

$$\|x\|^2 = \|Q^T x\|^2 + \|\tilde{Q}^T x\|^2$$

$\tilde{Q}$ dopełnienie ortoganalne macierzy Q.

MNK a rozkład QR

Jeśli $X = QR$ to

$$\|y - X\beta\|^2 = \|Q^T y - R\beta\|^2 + \|\tilde{Q}y\|^2$$

drugi składniki nie zależy od beta i dostajemy rozwiązanie MNK

$$Q^T y = R\beta.$$

Koszt rozwiązania wynosi $2(n - p/3)p^2$, czyli dla $n >> p$ koszt $2np^2$

Rozkład SVD

Dla dowolnej macierz $A \in \mathbb{R}^{n\times n}$ istnieje rozkład SVD, czyli

$$X = U\Sigma V^T,$$

U, V ortogonalne Σ diagonalana z wartościami
$\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_n \geqslant 0$.
Dla równania liniowego $Ax = b$ otrzymujemy

$$x = A^{-1}b = V\Sigma^{-1}U^T = \sum_{i=1}^{n} \frac{v_i u_i^T b}{\sigma_i}$$

Stabilność numeryczna równania liniowego

Rozpatrzmy problem

$$(\mathrm{A} + \varepsilon \mathrm{F})\mathrm{x}(\varepsilon) = \mathrm{b} + \varepsilon \mathrm{f}$$

Theorem

$$\frac{\|\mathrm{x}(\varepsilon) - \mathrm{x}\|}{\|\mathrm{x}\|} \leqslant \kappa(\mathrm{A})(\rho_{\mathrm{A}} + \rho_{\mathrm{b}}) + \mathcal{O}(\varepsilon^2)$$

gdzie $\kappa(\mathrm{A}) = \sigma_1(\mathrm{A})/\sigma_{\mathrm{n}}(\mathrm{A})$, $\rho_{\mathrm{A}} = |\varepsilon|\sigma_1(\mathrm{F})/\sigma_1(\mathrm{A})$, $\rho_{\mathrm{b}} = |\varepsilon|\|\mathrm{f}\|/\|\mathrm{b}\|$

Porównanie rozkładów QR i Choleskiego

Dla problemu

$$\min_{\beta} \|\mathrm{y} - \mathrm{X}\beta\|^2$$

- Korzystając z rozkładu Choleskiego wrażliwość rozwiązania jest rzędu $\kappa(\mathrm{X}^{\mathrm{T}}\mathrm{X})$.
- Korzystając z rozkładu QR nie korzystamy z macierzy $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ i można pokazać, że wrażliwość jest rzędu

$$\kappa(\mathrm{X}) + \kappa(\mathrm{X}^{\mathrm{T}}\mathrm{X}) * \|\mathrm{y} - \mathrm{X}\hat{\beta}\|.$$

Czyli rozkład Choleskiego jest tańszy ale bardziej wrażliwy na błedy numeryczne kiedy macierz planu jest źle uwarunkowana, w szczególnośći gdy $\|\mathrm{y} - \mathrm{X}\hat{\beta}\|$ jest małe.

Regresja logistyczna

$Y_1, Y_2, \ldots, Y_n$ - binarne zmienne objaśniane, $x_1, x_2, \ldots, x_n$ - zmienne objaśniające z $\mathbb{R}^p$. Oznaczmy przez $p_i = \mathbb{P}(Y_i = 1)$

Regresja logistyczna

$$\mathrm{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = x_i^T \beta$$

Log wiarygodność

$$\ell(\beta) = \sum_{i=1}^{n} Y_i x_i^T \beta - \sum_{i=1}^{n} \log(1 + \exp(x_i^T \beta))$$

Estymator największej wiarygodności

$$\hat{\beta} = \arg\max_{\beta} \ell(\beta) = \arg\min_{\beta} -\ell(\beta)$$

Jak obliczyć $\hat{\beta}$?

Algorytm Newtona - Raphsona

Niech f klasy C^2, rozwiązanie problemu

$$\arg\min_{x} f(x)$$

sprowadza się do rozwiązania równania

$$\nabla f(x) = 0 \tag{1}$$

Ze wzoru Taylora

$$\nabla f(x) \approx \nabla f(x_0) + \nabla^2 f(x_0)(x - x_0)$$

Czyli możemy przybliżać rozwiązanie (1) poprzez

$$x = x_0 - (\nabla^2 f(x_0))^{-1} \nabla f(x_0)$$

Algorytm Newtona-Raphsona

1. Ustalamy punkt startowy x_0.
2. Kolejne przybliżenia rozwiązania uzyskujemy

$$x_k = x_{k-1} - (\nabla^2 f(x_{k-1}))^{-1} \nabla f(x_{k-1})$$

Uogólnione modele liniowe

Regresja logistyczna jest szczególnym przypadkiem uogólnionych modeli liniowych. Rodzina wykładnicza

$$\mathrm{f}_\theta(\mathrm{y}) = \exp\left[\frac{\theta \mathrm{y} - \mathrm{b}(\theta)}{\mathrm{a}(\phi)}\right] \mathrm{h}(\mathrm{y})$$

- θ - parametr kanoniczny
- ϕ - parmetr dyspersji
- $\mathbb{E}(\mathrm{Y}) = \mathrm{b}'(\theta) = \mu$
- $\mathrm{var}(\mathrm{Y}) = \mathrm{a}(\phi)\mathrm{b}''(\theta)$

Uogólniony model liniowy $\mathrm{Y}_1, \ldots, \mathrm{Y}_\mathrm{n} \sim \mathrm{f}_{\theta_\mathrm{i}}$

$$\mathrm{g}(\mu_\mathrm{i}) = \mathrm{x}_\mathrm{i}^\mathrm{T}\beta = \eta_\mathrm{i}$$

g- funkcja łącząca (jeśli $\mathrm{g}(\mu_\mathrm{i}) = \theta_\mathrm{i}$ to kanoniczna funkcja łącząca)

Estymator największej wiarygodności w GLM I

$$\ell(\beta) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\theta_{\mathrm{i}}\mathrm{y} - \mathrm{b}(\theta_{\mathrm{i}})}{\mathrm{a}(\phi)}$$

Chcemy wyprowadzić algorytm Newtona-Raphsona (przybliżony).

$$\frac{\partial}{\partial \beta_{\mathrm{j}}}\ell = ? \quad \frac{\partial^2}{\partial \beta_{\mathrm{j}} \partial \beta_{\mathrm{k}}}\ell = ?$$

Dla uproszczenia zapisu przyjmijmy na chwilę $\mathrm{n} = 1$ i pomińmy indeks i

$$\frac{\partial}{\partial \beta_{\mathrm{j}}}\ell(\beta) = \frac{\partial \ell}{\partial \theta}\frac{\partial \theta}{\partial \mu}\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \beta_{\mathrm{j}}}$$

Estymator największej wiarygodności w GLM II

$$\frac{\partial \ell}{\partial \theta} = \frac{\mathrm{y} - \mathrm{b}'(\theta)}{\mathrm{a}(\phi)}$$
$$\frac{\partial \theta}{\partial \mu} = \frac{1}{\mathrm{b}''(\theta)} = \frac{\mathrm{a}(\phi)}{\mathrm{var}(\mathrm{Y})}$$
$$\frac{\partial \eta}{\partial \beta_{\mathrm{j}}} = \mathrm{x}_{\mathrm{ij}}$$

W przypadku kanonicznej funkcji łączącej $\eta = (\mathrm{b}')^{-1}(\mu)$

$$\frac{\partial \mu}{\partial \eta} = \mathrm{b}''(\theta)$$

Stąd

$$\frac{\partial \ell}{\partial \beta_{\mathrm{j}}} = \frac{\mathrm{y} - \mu}{\mathrm{var}(\mathrm{Y})} \mathrm{x}_{\mathrm{ij}} \frac{\partial \mu}{\partial \eta}$$

Estymator największej wiarygodności w GLM III

i w przypadku kanonicznej funnkcji łączącej

$$\frac{\partial \ell}{\partial \beta_{\mathrm{j}}} = \frac{\mathrm{y} - \mu}{\mathrm{a}(\phi)} \mathrm{x}_{\mathrm{ij}}$$

Fisher scoring

W algorytmie Newtona-Raphsona chcemy przybliżyć $\frac{\partial^2 \ell}{\partial \beta_{\mathrm{j}} \partial \beta_{\mathrm{k}}}$ przez $\mathbb{E} \frac{\partial^2}{\partial \beta_{\mathrm{j}} \partial \beta_{\mathrm{k}}} \ell$. Czyli

$$\nabla^2 \ell(\beta) \approx -\mathrm{I}(\beta)$$

$\mathrm{I}(\beta)$ - informacja Fishera

W przypadku kanonicznej funkcji łączacej

$$\nabla^2 \ell(\beta) = -\mathrm{I}(\beta)$$

Estymator największej wiarygodności w GLM IV

Korzystając z faktu

$$\mathbb{E}\frac{\partial^2 \ell}{\partial \beta_{\mathrm{j}} \partial \beta_{\mathrm{k}}} = -\mathbb{E}\frac{\partial \ell}{\partial \beta_{\mathrm{j}}}\frac{\partial \ell}{\partial \beta_{\mathrm{k}}}$$
$$= \mathbb{E}\frac{(\mathrm{y}-\mu)^2}{\mathrm{var}(\mathrm{Y})^2}\mathrm{x_{ij}x_{ik}}\left(\frac{\partial \mu}{\partial \eta}\right)^2 = \frac{\mathrm{x_{ij}x_{ik}}}{\mathrm{var}(\mathrm{Y})}\left(\frac{\partial \mu}{\partial \eta}\right)^2$$

Oznaczmy

$$\mathrm{A} = \mathrm{diag}\left(\frac{\frac{\partial \mu_{\mathrm{i}}}{\partial \eta_{\mathrm{i}}}}{\mathrm{var}(\mathrm{Y_i})}\right) \quad \mathrm{W} = \mathrm{diag}\left(\frac{\left(\frac{\partial \mu_{\mathrm{i}}}{\partial \eta_{\mathrm{i}}}\right)^2}{\mathrm{var}(\mathrm{Y_i})}\right)$$

Otrzymujemy

$$\nabla \ell(\beta) = \mathrm{X^T A}(\mathrm{y}-\mu), \quad \mathrm{I}(\beta) = \mathrm{X^T W X}$$

Estymator największej wiarygodności w GLM V

Oraz algorytm Fisher scoring przyjmuje postać

$$\begin{aligned}\beta^{k+1} &= \beta^{k} + (X^T W X)^{-1} X^T A (y - \mu) \\ &= (X^T W X)^{-1} [X^T W X \beta^{k} + X^T A (y - \mu)]\end{aligned}$$

$$A \operatorname{diag}\left(\frac{\partial \mu}{\partial \eta}\right) = W, \quad X^T \beta = \eta$$

Otrzymujemy iterowane ważone najmniejsze kwadraty

$$\beta^{k+1} = (X^T W^k X)^{-1} X^T W^k z^k,$$

gdzie

$$z = \eta - \operatorname{diag}\left(\frac{\partial \eta}{\partial \mu}\right)(y - \mu)$$

Powrót do regresji logistycznej

$Y_i \sim \mathrm{Bin}(1, p_i)$

$$\mathrm{logit}(p_i) = x_i^T \beta, \quad p_i = \frac{1}{1 + \exp(-x_i^t \beta)}$$

Gęstość można zapisać w postaci wykładniczej

$$f(y; p) = \exp(y\,\mathrm{logit}(p) + \log(1 - p)))$$

kanoniczny parametr $\theta = \mathrm{logit}(p)$

$$f(y; \theta) = \exp(y\theta - \log(1 + \exp(\theta)))$$

$$\mathbb{E}(Y) = p, \ \mathrm{var}(Y) = p(1 - p), \ \eta = \theta, \ \frac{\partial \eta}{\partial \mu} = \frac{1}{p(1 - p)}$$

Stąd W i z są postaci

$$W = \mathrm{diag}(p_i(1 - p_i)), \quad z_i = x_i^T \beta + \frac{y_i - p_i}{p_i(1 - p_i)}$$