MSiD

Zagadnienia wybrane.

JS

Lista 1

Zad. 1

Pokazać, że następujące tożsamości są prawdziwe:

a) $x^T x = \sum_i x_i^2$

b) $x^T A x = \sum_i \sum_j a_{ij} x_i x_j$

c) $tr(A^T X) = \sum_i \sum_j a_{ij} x_{ij}$

d) $A A^T = \sum_i a_i a_j^T$

Definicja mnożenia macierzy:

$A \cdot B = C$

$c_{ij} = a_{i1} b_{1j} + a_{i2} b_{2j} + \ldots + a_{in} b_{nj} = \sum_n a_{in} b_{nj}$

a)

$x^T x = x_{11}^T x_{11} + x_{12}^T x_{21} + x_{13}^T x_{31} + \ldots + x_{1n}^T x_{n1} = x_{11} x_{11} + x_{21} x_{21} + x_{31} x_{31} + \ldots + x_{n1} x_{n1} = \sum_i x_i^2$

b)

$$x^T A x = B x = C$$

$$c_{ij} = \sum_n b_{in} x_{nj} = \sum_n \sum_p x_{ip}^T a_{pn} x_{nj} = \sum_n \sum_p a_{pn} x_p x_n$$

c)

Definicja funkcji trace:

$tr(A) = \sum_n a_{nn}$

$$tr(A^T X) = tr(B) = \sum_n b_{nn} = \sum_n \sum_p a_{np}^T x_{pn} = \sum_n \sum_p a_{pn} x_{pn}$$

d)

Zad 2.

Policz gradienty:

a) $\nabla_x a^T x = a$

b) $\nabla_x x^T a = a$

c) $\nabla_x x^T A x = 2Ax$

d) $\nabla_x x^T A x = (A + A^T)x$

e) $\nabla_a log(a^T x) = \frac{1}{a^T x} x$

f) $\nabla_a \sigma(a^T x) = \sigma(a^T x)(1 - \sigma(a^T x))x$

g) $\nabla_A x^T A x = x x^T$

h) $\nabla_A \sigma(b^T \sigma(Ax))$

Dla funkcji $f(x_1, x_2, \cdots, x_n)$ gradient zapisujemy jako:
$\nabla f(x_1, x_2, \cdots, x_n) = [\frac{df}{dx_1}, \frac{df}{dx_2}, \cdots, \frac{df}{dx_n}]^T$.
A jest macierzą symetryczną, co oznacza $A^T = A$.

a)

Aby udowodnić równanie, obliczamy gradient funkcji dla jednej zmiennej:

$$\nabla_{x_k} a^T x = \frac{d}{dx_k} \sum_n a_n^T x_n =$$

Pochodna sumy równa się sumie pochodnych.

$$= \sum_n \frac{d}{dx_k} a_n^T x_n = \frac{d}{dx_k} a_1^T x_1 + \frac{d}{dx_k} a_2^T x_2 + \cdots + \frac{d}{dx_k} a_k^T x_k + \cdots + \frac{d}{dx_k} a_N^T x_N =$$

Pochodna ze stałej, czyli składniku nie zawierającego x_k, równa się zero.

$$= 0 + 0 + \cdots + \frac{d}{dx_k} a_k^T x_k + \cdots + 0 =$$

Obliczamy otrzymaną pochodną.

$$\frac{d}{dx_k} a_k^T x_k = a_k^T \frac{d}{dx_k} x_k = a_k^T$$

Dla każdego x_k z wektora **x** otrzymujemy pochodną a_k, z czego wynika, że dla wektora **x** otrzymamy wektor **a**.

b)

$$\nabla_{x_k} x^T a = \frac{d}{dx_k} \sum_n x_n^T a_n = \sum_n \frac{d}{dx_k} x^T a = 0 + \cdots + \frac{d}{dx_k} x_k^T a_k + \cdots + 0 = \frac{d}{dx_k} x_k^T a_k = a_k \frac{d}{dx_k} x_k^T = a_k$$

c)

$$\nabla_{x_k} x^T A x = \frac{d}{dx_k} \sum_i \sum_j a_{ij} x_i x_j = \sum_i \sum_j \frac{d}{dx_k} a_{ij} x_i x_j =$$

Bierzemy tylko te elementu z obu sum, które zawierają w sobie poszukiwany przez nas element x_k.

$$= 0 + \cdots + \sum_j \frac{d}{dx_k} a_{kj} x_k x_j + \cdots + \sum_i \frac{d}{dx_k} a_{ik} x_i x_k + \cdots + 0 =$$

Obliczamy otrzymane pochodne.

$$= \sum_j a_{kj} x_j + \sum_i a_{ik} x_i =$$

Przekształcamy otrzymane wyrażenie. Znak $\cdot$ na danej pozycji, oznacza wzięcie wszystkich elementów z tej pozycji, np. $b_{1.}$ oznacza wzięcie pierwszego wiersza z macierzy B.

$$= a_{k.} x_{.} + a_{.k} x_{.} = a_{k.} x_{.} + a_{k.}^T x_{.} =$$

Korzystając z własności macierzy symetrycznej otrzymujemy.

$$= a_{k.} x_{.} + a_{k.} x_{.} = 2 a_{k.} x_{.}$$

Otrzymane wyrażenie możemy odczytać jako, podwojony iloczyn k-tego wiersza macierzy A i wektora **x**. Dla wszystkich k oznacza to podwojony iloczyn macierzy A i wektora **x**.

d)

$$\nabla_{x_k} x^T A x = \frac{d}{dx_k} \sum_i \sum_j a_{ij} x_i x_j = \sum_i \sum_j \frac{d}{dx_k} a_{ij} x_i x_j =$$

Bierzemy tylko te elementu z obu sum, które zawierają w sobie poszukiwany przez nas element x_k.

$$= 0 + \cdots + \sum_j \frac{d}{dx_k} a_{kj} x_k x_j + \cdots + \sum_i \frac{d}{dx_k} a_{ik} x_i x_k + \cdots + 0 =$$

Obliczamy otrzymane pochodne.

$$= \sum_j a_{kj} x_j + \sum_i a_{ik} x_i =$$

Przekształcamy otrzymane wyrażenie. Znak $\cdot$ na danej pozycji, oznacza wzięcie wszystkich elementów z tej pozycji, np. $b_{1.}$ oznacza wzięcie pierwszego wiersza z macierzy B.

$$= a_{k.} x_{.} + a_{.k} x_{.} = a_{k.} x_{.} + a_{k.}^T x_{.} =$$

Po wyłączeniu wyrazów za nawias.

$$= (a_{k.} + a_{k.}^T) x_{.}$$

e)

$$\nabla_a log(a^T x) = \nabla_a log(\sum_i a_i^T x_i) = \frac{d}{da_k} log(\sum_i a_i^T x_i) =$$

Korzystamy ze wzoru na pochodną logarytmu.

$$= \frac{1}{\sum_i a_i^T x_i} \cdot \frac{d}{da_k} \sum_i a_i^T x_i =$$

Po wyliczeniu pochodnej otrzymujemy.

$$= \frac{1}{\sum_i a_i^T x_i} \cdot x_k$$

f)

$$\nabla_a \sigma(a^T x) = \frac{d}{da_k} \sigma(a^T x) =$$

Wzór na pochodną funkcji sigmoidalnej

$$\frac{d}{dx} \sigma(x) = \sigma(x) \cdot (1 - \sigma(x))$$

Funkcja wewnątrz sigmoidy jest funkcją złożoną, korzystając z twierdzenia o pochodnej z funkcji złożonych.

$$= \sigma(a^T x)(1 - \sigma(a^T x)) \cdot \frac{d}{da_k} \sum_i a_i^T x_i =$$

Po wyliczeniu pochodnej otrzymujemy.

$$= \sigma(a^T x)(1 - \sigma(a^T x))x_k$$

g)

$$\nabla_A x^T A x =$$

W tym przykładzie gradient wyliczamy po macierzy, chcąc obliczyć pochodną cząstkową, otrzymujemy

$$= \frac{d}{da_{kl}} x^T A x =$$

Oznacza to, że z macierzy A wybieramy jeden element a_{kl}.

$$= \frac{d}{da_{kl}} \sum_i \sum_j a_{ij} x_i x_j =$$

Najpierw wybieramy te składniki pierwszej sumy, które zawierają k-te wiersze macierzy A.

$$= 0 + \cdots + \frac{d}{da_{kl}} \sum_j a_{kj} x_k x_j + \cdots + 0 =$$

Spośród tych elementów wybieramy te, które zawierają elementy a_{kl}.

$$= 0 + \cdots + \frac{d}{da_{kl}} a_{kl} x_k x_l + \cdots + 0 =$$

Po obliczeniu pochodnej otrzymujemy.

$$= x_k x_l$$

h)

Zad 3.

Pokazać z definicji, że następujące funkcje są normami lub iloczynami skalarnymi (A – macierz symetryczna i dodatnio określona):

a) $\| x \|_2 = (\sum_d x_d^2)^{\frac{1}{2}}$

b) $\| x \|_A = \sqrt{x^T A x}$

c) $\langle x, y \rangle = x^T y$

d) $\langle x, y \rangle = x^T A y$

Daną funkcję nazywamy normą, gdy spełnia następujące warunki:

1.$\| x \| = 0 \Leftrightarrow x = 0,$

2.$\| \alpha x \| = |\alpha| \, \| x \|$, gdzie $\alpha \in R$

3.$\| x + y \| \leqslant \| x \| + \| y \|$

Daną funkcję nazywamy iloczynem skalarnym, gdy spełnia następujące warunki:

1.$\langle x, x \rangle \geqslant 0$

2.$\langle x, y \rangle = \langle y, x \rangle$

3.$\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$

4.$\langle x + z, y \rangle = \langle x, y \rangle + \langle z, y \rangle$

Macierz A jest dodatnio określona, gdy dla każdego niezerowego wektora **x** zachodzi następująca własność

$x^T A x > 0$

a)

Aby pokazać, że ta funkcja jest normą, musimy udowodnić, że spełnia ona wszystkie warunki normy.

$$\| x \|_2 = 0$$

$$(\sum_d x_d^2)^{\frac{1}{2}} = 0$$

$$\sum_d x_d^2 = 0$$

$$x_1^2 + x_2^2 + \cdots + x_d^2 + \cdots + x_D^2 = 0 \Leftrightarrow \forall x_d \in x : x_d = 0$$

$$\| \alpha x \|_2 = |\alpha| \| x \|_2$$

$$(\sum_d (\alpha x_d)^2)^{\frac{1}{2}} =$$

Wyłączamy stałą przed sumę.

$$= (\alpha^2 \sum_d x_d^2)^{\frac{1}{2}} =$$

Pierwiastek drugiego stopnia kwadratu liczby daje nam wartość bezwzględną.

$$= |\alpha| (\sum_d x_d^2)^{\frac{1}{2}} = |\alpha| \| x \|_2$$

$$\| x + y \|_2 \leqslant \| x \|_2 + \| y \|_2$$

$$(\sum_d (x_d + y_d)^2)^{\frac{1}{2}} \leqslant (\sum_d x_d^2)^{\frac{1}{2}} + (\sum_d y_d^2)^{\frac{1}{2}}$$

$$(\sum_d (x_d^2 + 2x_d y_d + y_d^2))^{\frac{1}{2}} \leqslant (\sum_d x_d^2)^{\frac{1}{2}} + (\sum_d y_d^2)^{\frac{1}{2}}$$

Podnosimy obie strony do kwadratu.

$$\sum_d (x_d^2 + 2x_d y_d + y_d^2) \leqslant \sum_d x_d^2 + 2(\sum_d x_d^2)^{\frac{1}{2}} (\sum_d y_d^2)^{\frac{1}{2}} + \sum_d y_d^2$$

$$\sum_d x_d^2 + \sum_d 2x_d y_d + \sum_d y_d^2 \leqslant \sum_d x_d^2 + 2(\sum_d x_d^2)^{\frac{1}{2}} (\sum_d y_d^2)^{\frac{1}{2}} + \sum_d y_d^2$$

Usuwamy wyrazy powtarzające się po obu stronach nierówności.

$$\sum_d 2x_d y_d \leqslant 2(\sum_d x_d^2)^{\frac{1}{2}} (\sum_d y_d^2)^{\frac{1}{2}}$$

Aby rozwiązać powyższe wyrażenie należy posłużyć się nierównością Cauchy'ego-Schwarza, nie jest to jednak wymagane.

b)

$$\| x \|_A = 0$$
$$\sqrt{x^T A x} = 0$$
Obie strony równania podnosimy do potęgi drugiej.
$$x^T A x = 0$$
Wynika to z własności macierzy dodatnio określonej.

$$\| \alpha x \|_A = |\alpha| \; \| x \|_A$$
$$\sqrt{(\alpha x)^T A \alpha x} = \sqrt{\alpha^2 x^T A x} = |\alpha| \sqrt{x^T A \alpha x}$$

$$\| x + y \|_A \leqslant \| x \|_A + \| y \|_A$$
$$\sqrt{(x + y)^T A (x + y)} \leqslant \sqrt{x^T A x} + \sqrt{y^T A y}$$
Korzystamy z własności operacji transponowania.
$$\sqrt{(x^T + y^T) A (x + y)} \leqslant \sqrt{x^T A x} + \sqrt{y^T A y}$$
Obie strony podnosimy do kwadratu.
$$(x^T + y^T) A (x + y) \leqslant x^T A x + 2\sqrt{x^T A x}\sqrt{y^T A y} + y^T A y$$
$$x^T A x + x^T A y + y^T A x + y^T A y \leqslant x^T A x + 2\sqrt{x^T A x}\sqrt{y^T A y} + y^T A y$$
Usuwamy wyrazy powtarzające się po obu stronach równania.
$$x^T A y + y^T A x + \leqslant 2\sqrt{x^T A x}\sqrt{y^T A y}$$
Korzystając z twierdzenia
$$(AB)^T = B^T A^T$$
dokonujemy przekształcenia
$$y^T A x = (y^T A x)^T = x^T (y^T A)^T = x^T A^T y = x^T A y$$
i podstawiamy do nierówności.
$$2x^T A y \leqslant 2\sqrt{x^T A x}\sqrt{y^T A y}$$
Obie strony równania dzielimy przez dwa i potęgujemy.
$$(x^T A y)^2 \leqslant x^T A x y^T A y$$
Dalsze udowadnianie nie jest wymagane.

c)

Aby pokazać, że ta funkcja jest iloczynem skalarnym, musimy udowodnić, że spełnia ona wszystkie warunki iloczynu skalarnego.

$$\langle x, x\rangle \geqslant 0$$
$$x^T x \geqslant 0$$
Co jest zawsze prawdą, ponieważ
$$\sum_n x_n^2 \geqslant 0$$

$$\langle x, y\rangle = \langle y, x\rangle$$
$$x^T y =$$
Korzystamy z twierdzenia
$$(AB)^T = B^T A^T.$$
$$(x^T y)^T = y^T x = \langle y, x\rangle$$

$$\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$$
$$(\alpha x)^T y = \alpha x^T y = \alpha\langle x, y\rangle$$

$$\langle x + z, y\rangle = \langle x, y\rangle + \langle z, y\rangle$$
$$(x + z)^T y =$$
Korzystamy z własności operacji transponowania.
$$= (x^T + z^T)y = x^T y + z^T y = \langle x, y\rangle + \langle z, y\rangle$$

d)

$$\langle x, x\rangle \geqslant 0$$
$$x^T A x = 0$$
Wynika to z własności macierzy dodatnio określonej.

$$\langle x, y\rangle = \langle y, x\rangle$$
$$x^T A y =$$
Korzystając z twierdzenia
$$(AB)^T = B^T A^T.$$
$$(x^T A y)^T = y^T (x^T A)^T = y^T A^T x = y^T A x = \langle y, x\rangle$$

$$\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$$
$$(\alpha x)^T A y = \alpha x^T A y = \alpha\langle x, y\rangle$$

$$\langle x + z, y\rangle = \langle x, y\rangle + \langle z, y\rangle$$
$$(x + z)^T A y =$$
Korzystamy z własności operacji transponowania.
$$(x^T + z^T)A y = x^T A y + z^T A y = \langle x, y\rangle + \langle z, y\rangle$$

Lista 2

Zad 1.
Pojawianie się spamu opisane jest zmienną losową x o rozkładzie dwupunktowym z parametrem $\theta \in [0,1]$, gdzie zmienna x przyjmuje wartość 1, jeśli pojawiająca się wiadomość jest spamem. Pewien użytkownik otagował N wiadomości. Korzystając z metody największej wiarygodności wyznaczyć estymator parametru θ.

Funkcja wiarygodności:
$p(D\,|\,\theta) = \prod_n p(x_n\,|\,\theta)$

Rozkład dwupunktowy:
$B(x\,|\,\theta) = \theta^x(1-\theta)^{1-x}$

Aby wyznaczyć estymator największej wiarygodności najpierw wyznaczamy funkcję wiarygodności dla wszystkich obserwacji a następnie znaleźć minimum tej funkcji.

$$p(D\,|\,\theta) = \prod_n p(x_n\,|\,\theta) = \prod_n B(x_n\,|\,\theta) = \prod_n \theta^{x_n}(1-\theta)^{1-x_n}$$

Tak otrzymaną funkcję logarytmujemy.

$$\log p(D\,|\,\theta) = \log \prod_n p(x_n\,|\,\theta) =$$

Korzystamy z własności logarytmów o iloczynie liczby logarytmowanej.

$$= \sum_n \log p(x_n\,|\,\theta) =$$

Dokonujemy kolejnych podstawień.

$$= \sum_n \log B(x_n\,|\,\theta) = \sum_n \log \theta^{x_n}(1-\theta)^{1-x_n} =$$

Korzystamy z własności logarytmu o potędze liczby logarytmowanej.

$$= \sum_n (x_n \log \theta + (1-x_n)\log(1-\theta))$$

Wyznaczenie estymatora największej wiarygodności sprowadza się teraz do znalezienia ekstremum minimalnego tej funkcji. W tym celu wyznaczymy pochodną tej funkcji po argumencie θ i przyrównamy ją do zera. Rozwiązaniem tego równania, będzie poszukiwany przez nas parametr.

$$\frac{d}{d\theta} \sum_n (x_n \log \theta + (1-x_n)\log(1-\theta)) = 0$$

Wyznaczamy pochodne.

$$\sum_n \left(\frac{x_n}{\theta} - \frac{1-x_n}{1-\theta}\right) = 0$$

$$\sum_n \left(\frac{x_n(1-\theta) - (1-x_n)\theta}{\theta(1-\theta)}\right) = 0$$

$$\sum_n (x_n(1-\theta) - (1-x_n)\theta) = 0$$

$$\sum_n (x_n - x_n\theta - \theta + x_n\theta) = 0$$

$$\sum_n (x_n - \theta) = 0$$

Wyłączamy element niezwiązany z sumą poza sumę.

$$\sum_n x_n - N\theta = 0$$

$$N\theta = \sum_n x_n$$

$$\theta_{ML} = \frac{\sum_n x_n}{N}$$

Zad 2.
Populacja studentów Politechniki Wrocławskiej została podzielona na trzy grupy:
1. Studenci osiągający średnią do 3.5.
2. Studenci osiągający średnią od 3.5 do 4.5.
3. Studenci osiągający średnią powyżej 4.5.
Populacja studentów opisana jest wektorem losowym $x = (x^1, x^2, x^3)^T$, przyjmującym trzy wartości $(1,0,0)^T$, gdy student należy do pierwszej grupy, $(0,1,0)^T$, gdy student należy do drugiej grupy i $(0,0,1)^T$, gdy student należy do trzeciej grupy. Rozkład zmiennej x wyraża się za pomocą rozkładu wielopunktowego o wektorze parametrów $\theta = (\theta_1, \theta_2, \theta_3)^T$. Z populacji studentów wybrano N obserwacji. Korzystając z metody największej wiarygodności wyliczyć estymator parametrów θ.

Rozkład wielopunktowy:
$$M(x \mid \theta) = \prod_d \theta_d^{x_d}$$

Wyznaczamy funkcję wiarygodności.
$$p(D \mid \theta) = \prod_n p(x_n \mid \theta) = \prod_n M(x_n \mid \theta) =$$
W tym momencie należy zaznaczyć, że funkcja wiarygodności jest iloczynem wszystkich obserwacji. Pojedyncza obserwacja jest dla nas wektorem, stąd zapis.
$$= \prod_n \prod_d \theta_d^{x_{dn}}$$

Logarytmujemy funkcję wiarygodności.
$$\log p(D \mid \theta) =$$
Dokonujemy podstawienia.
$$\sum_n \log M(x_n \mid \theta) = \sum_n \log \prod_d \theta_d^{x_{dn}} = \sum_n \sum_d \log \theta_d^{x_{dn}} =$$
Korzystamy z własności logarytmu o potędze liczby logarytmowanej.
$$= \sum_n \sum_d x_{dn} \log \theta_d$$

Szukana θ jest wektorem parametrów, co oznacza, że składa się z θ_1, θ_2, θ_3. Zarówno θ_1 i θ_2 należy obliczyć z pochodnych, natomiast θ_3 ze wzoru
$$\theta_3 = 1 - (\theta_1 + \theta_2).$$
Po wielu próbach oraz konsultacjach nie udało mi się dojść do sensownych wyników, dla zainteresowanych próbą rozwiązania, garść rad: wpisać wartość θ już w logarytmie funkcji i stworzyć układ równań z uzyskanych pochodnych przyrównanych do zera.

Zad 3.
Alarm samochodowy uzależnia swoje działanie od czujnika badającego poziom ultradźwięków w kabinie. Czujnik przed rozpoczęciem działania wymaga kalibracji. Przyjęto, że pomiary dokonywane przez czujnik są realizacjami zmiennej losowej x o rozkładzie normalnym $N(x\,|\,\mu,\sigma^2)$. Dokonano N pomiarów, gdy w kabinie nie występował żaden ruch. Korzystając z metody największej wiary- godności wyznaczyć estymatory parametrów μ i σ^2.

Rozkład normalny:

$$N(x\,|\,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$$

Wyznaczamy funkcję wiarygodności:

$$p(D\,|\,\theta) = \prod_n N(x_n\,|\,\mu,\sigma^2) =$$

Dokonujemy podstawienia.

$$= \prod_n \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x_n-\mu)^2}{2\sigma^2}\}$$

Logarytmujemy funkcję wiarygodności.

$$\log p(D\,|\,\theta) =$$

Dokonujemy podstawienia.

$$= \sum_n \log N(x_n\,|\,\mu,\sigma^2) = \sum_n \log(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x_n-\mu)^2}{2\sigma^2}\}) =$$

Korzystamy z własności logarytmów o iloczynie liczby logarytmowanej.

$$= \sum_n (\log(\frac{1}{\sqrt{2\pi\sigma^2}}) + \log(\exp\{-\frac{(x_n-\mu)^2}{2\sigma^2}\})) =$$

$$= \sum_n (-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_n-\mu)^2}{2\sigma^2}\log e)$$

Obliczamy parametry μ i σ^2.

$$\frac{d}{d\mu}\log p(D\,|\,\theta) = 0$$

$$\frac{d}{d\mu}\log p(D\,|\,\theta) = \frac{d}{d\mu}\sum_n (-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_n-\mu)^2}{2\sigma^2}) =$$

$$= \sum_n (\frac{d}{d\mu}(-\frac{1}{2}\log(2\pi\sigma^2)) - \frac{d}{d\mu}(\frac{(x_n-\mu)^2}{2\sigma^2})) =$$

Wyliczamy pochodne.

$$= \sum_n -\frac{1}{2\sigma^2}\frac{d}{d\mu}(x_n-\mu)^2 =$$

$$= \sum_n -\frac{1}{2\sigma^2}\frac{d}{d\mu}(x_n^2 - 2x_n\mu + \mu^2) =$$

$$= \sum_n -\frac{-2x_n + 2\mu}{2\sigma^2} = 0$$

Po wyliczeniu otrzymujemy.

$$\mu_{ML} = \frac{\sum_n x_n}{N}$$

$$\frac{d}{d\sigma}\log p(D|\theta) = 0$$

$$\frac{d}{d\sigma}\log p(D|\theta) = \frac{d}{d\sigma}\sum_n(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_n-\mu)^2}{2\sigma^2}) =$$

Wyliczamy pochodne.

$$= \sum_n(-\frac{1}{2}\frac{d}{d\sigma}\log(2\pi\sigma^2) - \frac{d}{d\sigma}\frac{(x_n-\mu)^2}{2\sigma^2}) =$$

$$= \sum_n(-\frac{1}{2}\frac{1}{2\pi\sigma^2}4\pi\sigma + (x_n-\mu)^2\frac{1}{\sigma^3}) = \sum_n(-\frac{1}{\sigma} + \frac{x_n-\mu}{\sigma^3}) =$$

Sprowadzamy do wspólnego mianownika.

$$= \sum_n(\frac{-\sigma^3 + (x_n-\mu)^2\sigma}{\sigma^4}) = 0$$

Przyrównujemy do zera.

$$\sum_n(-\sigma^3 + (x_n-\mu)^2\sigma) = 0$$

$$-N\sigma^3 + \sum_n(x_n-\mu)^2\sigma = 0$$

$$N\sigma^3 = \sigma\sum_n(x_n-\mu)^2$$

$$N\sigma^2 = \sum_n(x_n-\mu)^2$$

$$\sigma^2_{ML} = \frac{\sum_n(x_n-\mu)^2}{N}$$

Zad 4.
Charakterystyka wybranego słowa wypowiadanego przez człowieka opisana jest wektorem losowym cech $x = (x_1, \cdots, x_D)^T$ przyjmującym wartości z wielowymiarowego rozkładu normalnego $N(x \mid \mu, \Sigma)$. Pobrano N próbek danego słowa wypowiadanego przez różne osoby. Korzystając z metody największej wiarygodności wyznaczyć estymatory μ i Σ.

Rozkład normalny wielopunktowy:

$$N(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\}$$

Wybrane pochodne:

1. $\frac{d}{dy}(x-y)^T A(x-y) = -2(x-y)$
2. $\frac{d}{dA}(x-y)^T A^{-1}(x-y) = -A^{-1}(x-y)(x-y)^T A^{-1}$
3. $\frac{d}{dA} \log |A| = A^{-1}$

Wyznaczamy funkcję wiarygodności:

$$p(D \mid \theta) = \prod_n N(x_n \mid \mu, \Sigma) =$$

Dokonujemy podstawienia.

$$p(D \mid \theta) = \prod_n \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\{-\frac{1}{2}(x_n-\mu)^T \Sigma^{-1}(x_n-\mu)\}$$

Logarytmujemy funkcję wiarygodności.

$$\log p(D \mid \theta) =$$

Dokonujemy podstawienia.

$$= \sum_n \log N(x_n \mid \mu, \sigma^2) = \sum_n \log(\frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\{-\frac{1}{2}(x_n-\mu)^T \Sigma^{-1}(x_n-\mu)\}) =$$

Korzystamy z własności logarytmów o iloczynie liczby logarytmowanej.

$$= \sum_n (\log \frac{1}{(2\pi)^{D/2}} + \log \frac{1}{|\Sigma|^{1/2}} + \log exp\{-\frac{1}{2}(x_n-\mu)^T \Sigma^{-1}(x_n-\mu)\}) =$$

$$= \sum_n (-\frac{D}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2}(x_n-\mu)^T \Sigma^{-1}(x_n-\mu))$$

Obliczamy parametry μ i Σ.

$$\frac{d}{d\mu}\log p(D|\theta) = 0$$

$$\frac{d}{d\mu}N(x|\mu,\Sigma) = \frac{d}{d\mu}\sum_n(-\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)) =$$

$$= \sum_n(\frac{d}{d\mu}(-\frac{D}{2}\log 2\pi) + \frac{d}{d\mu}(-\frac{1}{2}\log|\Sigma|) + \frac{d}{d\mu}(-\frac{1}{2}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu))) =$$

$$= \sum_n(0 + 0 - \frac{1}{2}\frac{d}{d\mu}((x_n-\mu)^T\Sigma^{-1}(x_n-\mu))) =$$

Korzystamy z podanego wzoru (1) na pochodną.

$$= \sum_n(-\frac{1}{2}(-2)\Sigma^{-1}(x_n-\mu)) =$$

$$= \sum_n(\Sigma^{-1}(x_n-\mu)) =$$

$$= \Sigma^{-1}\sum_n(x_n-\mu) = 0$$

Przyrównujemy wyrażenie do zera.

$$\Sigma^{-1}\sum_n(x_n-\mu) = 0$$

Mnożymy stronami oba wyrażenia przez Σ od lewej strony.

$$\sum_n(x_n-\mu) = 0$$

$$\mu_{ML} = \frac{\sum_n x_n}{N}$$

$$\frac{d}{d\Sigma}\log p(D|\theta) = 0$$

$$\frac{d}{d\Sigma}N(x|\mu,\Sigma) = \frac{d}{d\Sigma}\sum_n(-\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)) =$$

$$= \sum_n(\frac{d}{d\Sigma}(-\frac{D}{2}\log 2\pi) + \frac{d}{d\Sigma}(-\frac{1}{2}\log|\Sigma|) + \frac{d}{d\Sigma}(-\frac{1}{2}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu))) =$$

$$= \sum_n(0 - \frac{1}{2}\frac{d}{d\Sigma}(\log|\Sigma|) - \frac{1}{2}\frac{d}{d\Sigma}((x_n-\mu)^T\Sigma^{-1}(x_n-\mu))) =$$

Korzystamy z podanych wzoru (2, 3) na pochodną.

$$= \sum_n(-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(x_n-\mu)(x_n-\mu)^T\Sigma^{-1}) = 0$$

Przyrównujemy wyrażenie do zera.

$$\sum_n(-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(x_n-\mu)(x_n-\mu)^T\Sigma^{-1}) = 0$$

Obie strony nierówności mnożymy przez Σ z prawej strony.

$$\sum_n(-\frac{1}{2}\Sigma^{-1}\Sigma + \frac{1}{2}\Sigma^{-1}(x_n-\mu)(x_n-\mu)^T\Sigma^{-1}\Sigma) = 0$$

$$\sum_n(-\frac{1}{2} + \frac{1}{2}\Sigma^{-1}(x_n-\mu)(x_n-\mu)^T) = 0$$

Obie strony nierówności mnożymy przez Σ z lewej strony.

$$\sum_n(-\frac{1}{2} + \frac{1}{2}\Sigma^{-1}(x_n-\mu)(x_n-\mu)^T) = 0$$

$$\sum_n (\Sigma(-\frac{1}{2}) + \frac{1}{2}\Sigma\Sigma^{-1}(x_n - \mu)(x_n - \mu)^T) = 0$$

$$\sum_n (-\frac{1}{2}\Sigma + \frac{1}{2}(x_n - \mu)(x_n - \mu)^T) = 0$$

$$-\frac{N}{2}\Sigma + \frac{1}{2}\sum_n (x_n - \mu)(x_n - \mu)^T = 0$$

$$\frac{N}{2}\Sigma = \frac{1}{2}\sum_n (x_n - \mu)(x_n - \mu)^T$$

$$\Sigma_{ML} = \frac{1}{N}\sum_n (x_n - \mu)(x_n - \mu)^T$$

Zad 5.
Niech zmienna losowa x ∈ {0,1} oznacza odpowiednio porażkę lub zwycięstwo Śląska Wrocław w meczu. Zmienna x opisana jest rozkładem dwupunktowym B(x | θ). Zebrano wyniki N spotkań. Przyjmując rozkład a priori Beta(θ | a, b), wyznaczyć estymator MAP (maksymalnego a posteriori) parametru θ. Jak można zinterpretować parametry a i b?

Rozkład a posteriori:
$$p(\theta|D) = \frac{p(\theta)\prod_n p(x_n|\theta)}{p(D)}$$

Rozkład dwupunktowy:
$$B(x|\theta) = \theta^x(1-\theta)^{1-x}$$

Rozkład Beta:
$$Beta(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

Wyznaczamy rozkład a posteriori.
$$p(\theta|D) = \frac{p(\theta)\prod_n p(x_n|\theta)}{p(D)} =$$
Dokonujemy podstawienia.
$$p(\theta|D) = \frac{\frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\prod_n \theta^{x_n}(1-\theta)^{1-x_n}}{p(D)} =$$
Funkcję gamma z rozkładu beta traktujemy jako stałą.
$$= \frac{C\theta^{a-1}(1-\theta)^{b-1}\prod_n \theta^{x_n}(1-\theta)^{1-x_n}}{p(D)}$$

Logarytmujemy funkcję rozkładu.
$$\log p(\theta|D) =$$
Dokonujemy podstawienia.
$$= \log(\frac{C\theta^{a-1}(1-\theta)^{b-1}\prod_n \theta^{x_n}(1-\theta)^{1-x_n}}{p(D)}) =$$
$$= \log C + \log\theta^{a-1} + \log(1-\theta)^{b-1} + \log\prod_n \theta^{x_n}(1-\theta)^{1-x_n} - \log p(D) =$$
$$= \log C + (a-1)\log\theta + (b-1)\log(1-\theta) + \sum_n \log\theta^{x_n}(1-\theta)^{1-x_n} - \log p(D) =$$
$$= \log C + (a-1)\log\theta + (b-1)\log(1-\theta) + \sum_n (x_n\log\theta + (1-x_n)\log(1-\theta)) - \log p(D)$$

Odszukujemy minimum funkcji poprzez przyrównanie gradientu funkcji do zera.
$$\frac{d}{d\theta}\log p(\theta|D) = 0$$
$$\frac{d}{d\theta}\log p(\theta|D) = \frac{d}{d\theta}(\log C + (a-1)\log\theta + (b-1)\log(1-\theta) + \sum_n (x_n\log\theta + (1-x_n)\log(1-\theta)) - \log p(D)) =$$
$$= (\frac{d}{d\theta}\log C + \frac{d}{d\theta}(a-1)\log\theta + \frac{d}{d\theta}(b-1)\log(1-\theta) + \frac{d}{d\theta}\sum_n (x_n\log\theta + (1-x_n)\log(1-\theta)) - \frac{d}{d\theta}\log p(D)) =$$
$$= 0 + \frac{a-1}{\theta} - \frac{b-1}{1-\theta} + \sum_n (\frac{x_n}{\theta} - \frac{1-x_n}{1-\theta}) - 0 =$$

$$= \frac{a-1}{\theta} - \frac{b-1}{1-\theta} + \sum_n \frac{x_n}{\theta} - \sum_n \frac{1-x_n}{1-\theta} =$$
$$= \frac{a-1}{\theta} - \frac{b-1}{1-\theta} + \frac{1}{\theta}\sum_n x_n - \frac{1}{1-\theta}\sum_n (1-x_n) = 0$$

Sprowadzamy całe wyrażenie do wspólnego mianownika i przyrównujemy do zera.

$$(a-1)(1-\theta) - (b-1)\theta + (1-\theta)\sum_n x_n - \theta\sum_n (1-x_n) = 0$$
$$a - \theta a - 1 + \theta - \theta b + \theta + \sum_n x_n - \theta\sum_n x_n - \theta N + \theta\sum_n x_n = 0$$
$$-\theta(a + b - 2 + N) + a - 1 + \sum_n x_n = 0$$

Co ostatecznie daje nam wynik.

$$\theta_{MAP} = \frac{a - 1 + \sum_n x_n}{N + a + b - 2}$$

Kolokwium A MSiD

Zad 1.
Dany jest wektor cech $\phi = (\phi_1, \phi_2)$ oraz prawdopodobieństwa $p(\phi_1 | y = 0) = 0{,}4$, $p(\phi_1 | y = 1) = 0{,}2$, $p(\phi_2 | y = 0) = 0{,}7$, $p(\phi_2 | y = 1) = 0{,}5$ oraz $p(y = 0) = 0{,}4$. Zakładając, że $p(\phi_1, \phi_2 | y) = p(\phi_1 | y)p(\phi_2 | y)$ podać w karcie odpowiedzi:
1.1 Prawdopodobieństwo $p(y = 1)$
1.2 Prawdopodobieństwo $p(\phi | y = 0)$
1.3 Prawdopodobieństwo $p(y = 0 | \phi)$
1.4 Prawdopodobieństwo $p(y = 1 | \phi)$
1.5 Klasę, do której zaklasyfikujemy wektor ϕ

Twierdzenie Bayesa:
$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

1.1

Obliczamy prawdopodobieństwo etykiety y równej 1
$$p(y = 0) + p(y = 1) = 1$$
Wynika to z reguł prawdopodobieństwa.
$$p(y = 1) = 1 - p(y = 0)$$
Dokonujemy podstawienia.
$$p(y = 1) = 1 - 0{,}4$$
$$p(y = 1) = 0{,}6$$

1.2

Wyliczamy prawdopodobieństwo całej klasy pod warunkiem etykiety y
$$p(\phi | y = 0)$$
Podstawiamy wartości do podanego wzoru.
$$p(\phi | y = 0) = p(\phi_1 | y = 0)p(\phi_2 | y = 0)$$
$$p(\phi | y = 0) = 0{,}4 \cdot 0{,}7$$
$$p(\phi | y = 0) = 0{,}28$$

1.3

Obliczamy jakie jest prawdopodobieństwo etykiety y równej 0 pod warunkiem całej klasy ϕ
$$p(y = 0 | \phi)$$
Z twierdzenia Bayesa.
$$p(\phi | y = 0) = \frac{p(y = 0 | \phi)p(\phi)}{p(y = 0)}$$
Po przekształceniu otrzymujemy.
$$p(y = 0 | \phi) = \frac{p(\phi | y = 0)p(y = 0)}{p(\phi)}$$
Znane nam są $p(\phi | y = 0)$ oraz $p(y = 0)$, musimy obliczyć prawdopodobieństwo całej klasy $p(\phi)$.
$$p(\phi) = p(\phi_1 | y = 0)p(\phi_2 | y = 0)p(y = 0) + p(\phi_1 | y = 1)p(\phi_2 | y = 1)p(y = 1) = 0{,}4 \cdot 0{,}7 \cdot 0{,}4 + 0{,}2 \cdot 0{,}5 \cdot 0{,}6 \approx 0{,}172$$
Otrzymaną wartość wstawiamy do równania.
$$p(y = 0 | \phi) \approx \frac{0{,}28 \cdot 0{,}4}{0{,}172} \approx 0{,}651$$

1.4

Obliczyć prawdopodobieństwo etykiety y równej 1 pod warunkiem całej klasy możemy na dwa sposoby, analogicznie do poprzedniego podpunktu, wyliczyć prawdopodobieństwo klasy ϕ dla etykiety y równej 1 i skorzystać z twierdzenia Bayesa, lub ponownie wykorzystać własności rachunku prawdopodobieństwa.

$$1 = p(y=0\,|\,\phi) + p(y=1\,|\,\phi)$$

Z czego wynika.

$$p(y=1\,|\,\phi) = 1 - p(y=0\,|\,\phi)$$

Po podstawieniu.

$$p(y=1\,|\,\phi) \approx 1 - 0{,}651 \approx 0{,}349$$

1.5

Dokonujemy predykcji klasy ϕ. Wiedząc, że pod warunkiem klasy ϕ największe prawdopodobieństwo ma klasa y równa 0,

$$p(y=0\,|\,\phi) \approx 0{,}651$$

$$p(y=1\,|\,\phi) \approx 0{,}349\,,$$

naszą predykcją jest właśnie ta klasa.

Zad 2.

Dana jest zmienna losowa x ∈ R o rozkładzie normalnym oraz rozkład a priori na μ:

$$N(x\,|\,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\},$$

$$p(\mu) = \frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}\mu^2\}$$

gdzie $\mu \in \mathbb{R}$, $\sigma > 0$. Dla obserwacji $D = \{x_1, \cdots, x_N\}$ oraz **znanego** σ^2 podać w karcie odpowiedzi:

2.1 Estymator maksymalnego a posteriori (MAP) dla parametru μ.

2.2 Wartości estymatora dla $D = \{-1,1,2\}$ oraz $\sigma^2 = 1$.

Rozkład a posteriori:

$$p(\theta\,|\,D) = \frac{p(\theta)\prod_n p(x_n\,|\,\theta)}{p(D)}$$

2.1

Wyznaczamy rozkład a posteriori dla parametru μ.

$$p(\mu\,|\,D) =$$

Dokonujemy podstawień (zmieniam konwencję ku czytelności).

$$= \frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}\mu^2\}\cdot\prod_n\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x_n-\mu)^2}{2\sigma^2}\} \div p(D)$$

Logarytmujemy funkcję rozkładu.

$$\log p(\mu\,|\,D) =$$

$$= \log\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}\mu^2\} + \log\prod_n\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x_n-\mu)^2}{2\sigma^2}\} - \log p(D) =$$

Upraszczamy wyrażenie.

$$= \log\frac{1}{\sqrt{2\pi}} + \log\exp\{-\frac{1}{2}\mu^2\} + \sum_n\log\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x_n-\mu)^2}{2\sigma^2}\} - \log p(D) =$$

$$= \log\frac{1}{\sqrt{2\pi}} + \log\exp\{-\frac{1}{2}\mu^2\} + \sum_n(\log\frac{1}{\sqrt{2\pi\sigma^2}} + \log\exp\{-\frac{(x_n-\mu)^2}{2\sigma^2}\}) - \log p(D) =$$

$$= \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2}\mu^2 + N \log \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_n \frac{(x_n - \mu)^2}{2\sigma^2} - \log p(D) =$$
$$= \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2}\mu^2 + N \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_n (x_n - \mu)^2 - \log p(D)$$

Znalezienie poszukiwanego parametru sprowadza się teraz do znalezienia ekstremum tej funkcji i wyliczenie jej wartości w tym punkcie.

$$\frac{d}{d\mu} \log p(\mu \,|\, D) =$$

Podstawiamy wcześniejsze wyrażenie.

$$= \frac{d}{d\mu}(\log \frac{1}{\sqrt{2\pi}} - \frac{1}{2}\mu^2 + N \log \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_n \frac{(x_n - \mu)^2}{2\sigma^2} - \log p(D)) =$$
$$= 0 - \mu + 0 - \frac{1}{2\sigma^2} \sum_n \frac{d}{d\mu}(x_n - \mu)^2 - 0 =$$
$$= -\mu - \frac{1}{2\sigma^2} \sum_n \frac{d}{d\mu}(x_n^2 - 2x_n\mu + \mu^2) =$$
$$= -\mu - \frac{1}{2\sigma^2} \sum_n (-2x_n + 2\mu) =$$
$$= -\mu - \frac{1}{\sigma^2} \sum_n (-x_n + \mu)$$

Tak uproszczone wyrażenie przyrównujemy do zera.

$$-\mu - \frac{1}{\sigma^2} \sum_n (-x_n + \mu) = 0$$
$$-\mu + \frac{1}{\sigma^2} \sum_n x_n - \frac{N}{\sigma^2}\mu = 0$$
$$-\mu(1 + \frac{N}{\sigma^2}) + \frac{1}{\sigma^2} \sum_n x_n = 0$$
$$\mu(1 + \frac{N}{\sigma^2}) = \frac{1}{\sigma^2} \sum_n x_n$$
$$\mu(\frac{\sigma^2 + N}{\sigma^2}) = \frac{1}{\sigma^2} \sum_n x_n$$
$$\mu = \frac{1}{\sigma^2} \sum_n x_n \cdot \frac{\sigma^2}{\sigma^2 + N}$$
$$\mu = \frac{1}{\sigma^2 + N} \sum_n x_n$$

2.2

Obliczenie wartości estymatora dla podanych danych.

$$\mu = \frac{1}{\sigma^2 + N} \sum_n x_n =$$

Po wstawieniu danych.

$$= \frac{1}{1^2 + 3}(-1 + 1 + 2) = \frac{1}{2}$$

Zad 3.
Dane są wektory $x \in \mathbb{R}^M$, $y \in \mathbb{R}^M$, macierz $A \in \mathbb{R}^{NxM}$ oraz funkcja

$$f(A) = x^T A^T A y.$$

Podać w karcie odpowiedzi:

3.1. Pochodną cząstkową $\dfrac{d}{d a_{ij}}$

3.2. Gradient $\nabla_A f(A)$ zapisany bez użycia jakiejkolwiek sumy.

Kolokwium C MSiD

Zad 1.
Dany jest wektor cech $\phi = (\phi_1, \phi_2)$ oraz prawdopodobieństwa $p(\phi_1 | y = 0) = 0{,}6$, $p(\phi_1 | y = 1) = 0{,}5$, $p(\phi_2 | y = 0) = 0{,}4$, $p(\phi_2 | y = 1) = 0{,}3$ oraz $p(y = 0) = 0{,}6$. Zakładając, że $p(\phi_1, \phi_2 | y) = p(\phi_1 | y)p(\phi_2 | y)$ podać w karcie odpowiedzi:
1.1 Prawdopodobieństwo $p(y = 1)$
1.2 Prawdopodobieństwo $p(\phi | y = 0)$
1.3 Prawdopodobieństwo $p(y = 0 | \phi)$
1.4 Prawdopodobieństwo $p(y = 1 | \phi)$
1.5 Klasę, do której zaklasyfikujemy wektor ϕ

Twierdzenie Bayesa:
$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$

1.1

Obliczamy prawdopodobieństwo etykiety y równej 1
$$p(y = 0) + p(y = 1) = 1$$
Wynika to z reguł prawdopodobieństwa.
$$p(y = 1) = 1 - p(y = 0)$$
Dokonujemy podstawienia.
$$p(y = 1) = 1 - 0{,}6$$
$$p(y = 1) = 0{,}4$$

1.2

Wyliczamy prawdopodobieństwo całej klasy pod warunkiem etykiety y
$$p(\phi | y = 0)$$
Podstawiamy wartości do podanego wzoru.
$$p(\phi | y = 0) = p(\phi_1 | y = 0)p(\phi_2 | y = 0)$$
$$p(\phi | y = 0) = 0{,}6 \cdot 0{,}4$$
$$p(\phi | y = 0) = 0{,}24$$

1.3

Obliczamy jakie jest prawdopodobieństwo etykiety y równej 0 pod warunkiem całej klasy ϕ
$$p(y = 0 | \phi)$$
Z twierdzenia Bayesa.
$$p(\phi | y = 0) = \frac{p(y = 0 | \phi)p(\phi)}{p(y = 0)}$$
Po przekształceniu otrzymujemy.
$$p(y = 0 | \phi) = \frac{p(\phi | y = 0)p(y = 0)}{p(\phi)}$$
Znane nam są $p(\phi | y = 0)$ oraz $p(y = 0)$, musimy obliczyć prawdopodobieństwo całej klasy $p(\phi)$.
$$p(\phi) = p(\phi_1 | y = 0)p(\phi_2 | y = 0)p(y = 0) + p(\phi_1 | y = 1)p(\phi_2 | y = 1)p(y = 1) = 0{,}6 \cdot 0{,}4 \cdot 0{,}6 + 0{,}5 \cdot 0{,}3 \cdot 0{,}4 \approx 0{,}204$$
Otrzymaną wartość wstawiamy do równania.
$$p(y = 0 | \phi) \approx \frac{0{,}24 \cdot 0{,}6}{0{,}204} \approx 0{,}706$$

1.4

Obliczyć prawdopodobieństwo etykiety y równej 1 pod warunkiem całej klasy możemy na dwa sposoby, analogicznie do poprzedniego podpunktu, wyliczyć prawdopodobieństwo klasy ϕ dla etykiety y równej 1 i skorzystać z twierdzenia Bayesa, lub ponownie wykorzystać własności rachunku prawdopodobieństwa.

$$1 = p(y=0\,|\,\phi) + p(y=1\,|\,\phi)$$

Z czego wynika.

$$p(y=1\,|\,\phi) = 1 - p(y=0\,|\,\phi)$$

Po podstawieniu.

$$p(y=1\,|\,\phi) \approx 1 - 0{,}706 \approx 0{,}294$$

1.5

Dokonujemy predykcji klasy ϕ. Wiedząc, że pod warunkiem klasy ϕ największe prawdopodobieństwo ma klasa y równa 0,

$$p(y=0\,|\,\phi) \approx 0{,}706$$

$$p(y=1\,|\,\phi) \approx 0{,}294\,,$$

naszą predykcją jest właśnie ta klasa.

Zad 2.

Dana jest zmienna losowa x ∈ R o rozkładzie normalnym oraz rozkład a priori na μ:

$$N(x\,|\,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x-\log\mu)^2}{2\sigma^2}\},$$

$$p(\mu) = \frac{1}{\mu\sqrt{2\pi}}\exp\{-\frac{1}{2}(\log\mu)^2\}$$

gdzie $\mu\in\mathbb{R}$, $\sigma>0$. Dla obserwacji $D=\{x_1,\cdots,x_N\}$ oraz **znanego** σ^2 podać w karcie odpowiedzi:

2.1 Estymator maksymalnego a posteriori (MAP) dla parametru μ.

2.2 Wartości estymatora dla $D=\{-1,1,2\}$ oraz $\sigma^2=1$.

Rozkład a posteriori:

$$p(\theta\,|\,D) = \frac{p(\theta)\prod_n p(x_n\,|\,\theta)}{p(D)}$$

2.1

Wyznaczamy rozkład a posteriori dla parametru μ.

$$p(\mu\,|\,D) =$$

Dokonujemy podstawień (zmieniam konwencję ku czytelności).

$$= \frac{1}{\mu\sqrt{2\pi}}\exp\{-\frac{1}{2}(\log\mu)^2\}\cdot\prod_n\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x_n-\log\mu)^2}{2\sigma^2}\}\div p(D)$$

Logarytmujemy funkcję rozkładu.

$$\log p(\mu\,|\,D) =$$

$$= \log\frac{1}{\mu\sqrt{2\pi}}\exp\{-\frac{1}{2}(\log\mu)^2\} + \log\prod_n\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x_n-\log\mu)^2}{2\sigma^2}\} - \log p(D) =$$

Upraszczamy wyrażenie.

$$= \log\frac{1}{\mu\sqrt{2\pi}} + \log\exp\{-\frac{1}{2}(\log\mu)^2\} + \sum_n\log\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x_n-\log\mu)^2}{2\sigma^2}\} - \log p(D) =$$

$$= \log\frac{1}{\mu\sqrt{2\pi}} + \log\exp\{-\frac{1}{2}(\log\mu)^2\} + \sum_n(\log\frac{1}{\sqrt{2\pi\sigma^2}} + \log\exp\{-\frac{(x_n-\log\mu)^2}{2\sigma^2}\}) - \log p(D) =$$

$$= \log \frac{1}{\mu\sqrt{2\pi}} - \frac{1}{2}(\log \mu)^2 + N \log \frac{1}{\sqrt{2\pi\sigma^2}} - \sum_n \frac{(x_n - \log \mu)^2}{2\sigma^2} - \log p(D) =$$
$$= \log \frac{1}{\mu} + \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2}(\log \mu)^2 + N \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_n (x_n - \log \mu)^2 - \log p(D)$$

Znalezienie poszukiwanego parametru sprowadza się teraz do znalezienia ekstremum tej funkcji i wyliczenie jej wartości w tym punkcie.

$$\frac{d}{d\mu} \log p(\mu \,|\, D) =$$

Podstawiamy wcześniejsze wyrażenie.

$$= \frac{d}{d\mu}(\log \frac{1}{\mu} + \log \frac{1}{\sqrt{2\pi}} - \frac{1}{2}(\log \mu)^2 + N \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_n (x_n - \log \mu)^2 - \log p(D)) =$$
$$= -\frac{1}{\mu} - \frac{\log \mu}{\mu} + 0 - \frac{1}{2\sigma^2} \sum_n \frac{d}{d\mu}(x_n - \log \mu)^2 - 0 =$$
$$= -\frac{1}{\mu} - \frac{\log \mu}{\mu} - \frac{1}{2\sigma^2} \sum_n \frac{d}{d\mu}(x_n^2 - 2x_n \log \mu + \log^2 \mu) =$$
$$= -\frac{1}{\mu} - \frac{\log \mu}{\mu} - \frac{1}{2\sigma^2} \sum_n (-\frac{2(x_n - \log \mu)}{\mu}) =$$
$$= -\frac{1}{\mu} - \frac{\log \mu}{\mu} + \frac{1}{\sigma^2} \sum_n \frac{(x_n - \log \mu)}{\mu}$$

Tak uproszczone wyrażenie przyrównujemy do zera.

$$-\frac{1}{\mu} - \frac{\log \mu}{\mu} + \frac{1}{\sigma^2} \sum_n \frac{(x_n - \log \mu)}{\mu} = 0$$
$$-\frac{1}{\mu} - \frac{\log \mu}{\mu} + \frac{1}{\mu\sigma^2} \sum_n (x_n - \log \mu) = 0$$

Sprowadzamy całość do wspólnego mianownika.

$$-\frac{\sigma^2}{\mu\sigma^2} - \frac{\sigma^2 \log \mu}{\mu\sigma^2} + \frac{1}{\mu\sigma^2} \sum_n (x_n - \log \mu) = 0$$
$$-\sigma^2 - \sigma^2 \log \mu + \sum_n (x_n - \log \mu) = 0$$
$$-\sigma^2 - \sigma^2 \log \mu + \sum_n x_n - N \log \mu = 0$$
$$\sigma^2 \log \mu + N \log \mu = -\sigma^2 + \sum_n x_n$$
$$\log \mu(\sigma^2 + N) = -\sigma^2 + \sum_n x_n$$
$$\log \mu = \frac{-\sigma^2 + \sum_n x_n}{(\sigma^2 + N)}$$

Odwracamy funkcję.

$$\mu = \exp\{\frac{-\sigma^2 + \sum_n x_n}{(\sigma^2 + N)}\}$$

2.2

Obliczenie wartości estymatora dla podanych danych.

$$\mu = \exp\{\frac{-\sigma^2 + \sum_n x_n}{(\sigma^2 + N)}\} =$$

Po wstawieniu danych.

$$= \exp\{\frac{-2 + (-1 + 1 + 2)}{2 + 3}\} = \exp\frac{0}{5} = 1$$

Kolokwium C ROiS

Zad 1.
Dany jest model κ-najbliższych s¡siadów i ciąg uczący
$\{(-2x+2,1),(4x+2,0),(3,1),(-2x^2+0.5,0)\}$, zawieraj¡cy pary (f_n, y_n) gdzie f_n jest funkcją zmiennej x, a $y_n \in \{0,1\}$ numerem klasy. Wykorzystując odległość

$$d(f,g) = \max_{x\in[0,1]} |f(x) - g(x)|,$$

dla nowej obserwacji f = −2x podać w karcie odpowiedzi:
1.1. Odległość d(f, f1)
1.2. Odległość d(f, f4)
1.3. Prawdopodobieństwo (y = 1|f) dla κ = 1
1.4. Prawdopodobieństwo p(y = 1|f) dla κ = 3
1.5. Klasę, do której zaklasyfikujemy f dla κ = 3

1.1

Obliczamy odległość wykorzystując podany wzór.

$$d(f,g) = \max_{x\in[0,1]} |f(x) - g(x)| =$$

Dokonujemy podstawień.

$$= \max_{x\in[0,1]} |f - f_1| =$$

$$= \max_{x\in[0,1]} |-2x - (-2x+2)| =$$

$$= \max_{x\in[0,1]} |-2| =$$

Wartość maksymalna ze stałej funkcji wynosi.

$$= 2$$

1.2

Obliczamy odległość wykorzystując podany wzór.

$$d(f,g) = \max_{x\in[0,1]} |f(x) - g(x)| =$$

Dokonujemy podstawień.

$$= \max_{x\in[0,1]} |f - f_4| =$$

$$= \max_{x\in[0,1]} |-2x - (4x+2)| =$$

$$= \max_{x\in[0,1]} |-2x - (2x^2+0,5)| =$$

$$= \max_{x\in[0,1]} |2x^2 - 2x - 0,5|$$

Szukamy teraz największej wartości jaką przyjmuje ta funkcja w przedziale od 0 włącznie do 1 włącznie. Wykorzystując pochodne wiemy, że funkcja ta ma swoje ekstremum w punkcie $\frac{1}{2}$. Wartość ta należy do przedziału funkcji max. Po podstawieniu tej wartości otrzymujemy wynik.

$$= \max_{x\in[0,1]} |2(\frac{1}{2})^2 - 2\frac{1}{2} - 0,5| = 1$$

1.3

Obliczamy, jakie jest prawdopodobieństwo, że dla k najbliższych sąsiadów „wylosujemy" tego sąsiada, który ma etykietę y. Zacznijmy od obliczenia odległości wszystkich par (f_n, y_n) od naszej obserwacji f.

$$d(f, f_1) = 2$$
$$d(f, f_2) = 8$$
$$d(f, f_3) = 5$$
$$d(f, f_4) = 1$$

Z treści zadania wynika, że pod uwagę bierzemy jednego najbliższego sąsiada i sprawdzamy jakie jest prawdopodobieństwo, że etykieta „wylosowanego" sąsiada jest równa 1. Najbliższym sąsiadem jest obserwacja f_4. Jej etykieta y_n równa 0. Obliczamy prawdopodobieństwo.

$$\frac{0}{1} = 0$$

1.4

Postępujemy analogicznie do poprzedniego zadania, z tą różnicą, że teraz badamy prawdopodobieństwa klasy dla 3 najbliższych sąsiadów, którymi są następujące obserwacje.

$$(f_1, y_1 = 1)$$
$$(f_3, y_3 = 1)$$
$$(f_4, y_4 = 0)$$

Prawdopodobieństwo wynosi.

$$\frac{2}{3}$$

1.5

Ponieważ największe prawdopodobieństwo obserwacji f dla k=3 jest etykieta równa 1, taka też jest nasz predykcja.

Zad 2.
Dana jest zmienna losowa x > 0 o rozkładzie Gamma:

$$p(x\,|\,\lambda, k) = \frac{1}{\lambda^k \Gamma(k)} x^{k-1} \exp\{-\frac{x}{\lambda}\}$$

gdzie $\lambda, k > 0, \Gamma(\,\cdot\,)$ oznacza funkcję gamma. Dla obserwacji $D = \{x_1, \cdots, x_N\}$ oraz **znanego** k podać w karcie odpowiedzi:

2.1 Funkcję wiarygodności $p(D\,|\,\lambda)$

2.2 Estymator największej wiarygodności (ML) dla parametru λ

2.3 Wartości estymatora dla $D = \{1,1,2,4\}$ oraz $k = 0.5$

2.1

Funkcja wiarygodności.

$$p(D\,|\,\theta) = \prod_n p(x_n\,|\,\theta) =$$

Podstawiamy funkcję rozkładu.

$$= \prod_n p(x_n\,|\,\lambda, k) = \prod_n \frac{1}{\lambda^k \Gamma(k)} x_n^{k-1} \exp\{-\frac{x_n}{\lambda}\}$$

2.2

Obliczamy estymator największej wiarygodności dla parametru λ, zaczynając od zlogarytmowania funkcji wiarygodności.

$$p(D\,|\,\theta) = \log \prod_n \frac{1}{\lambda^k \Gamma(k)} x_n^{k-1} \exp\{-\frac{x_n}{\lambda}\} =$$

$$= \sum_n \log \frac{1}{\lambda^k \Gamma(k)} x_n^{k-1} \exp\{-\frac{x_n}{\lambda}\} =$$

$$= \sum_n (\log \frac{1}{\lambda^k} + \log \frac{1}{\Gamma(k)} + \log x_n^{k-1} + \log \exp\{-\frac{x_n}{\lambda}\}) =$$

$$= \sum_n (-k \log \lambda - \log \Gamma(k) + (k-1) \log x_n - \frac{x_n}{\lambda})$$

Z tak zlogarytmowanego wyrażenia obliczamy teraz gradient po interesującym nas estymatorze, tj. λ i przyrównujemy do zera.

$$\frac{d}{d\lambda} \sum_n (-k \log \lambda - \log \Gamma(k) + (k-1) \log x_n - \frac{x_n}{\lambda}) =$$

$$= \sum_n (-\frac{d}{d\lambda} k \log \lambda - \frac{d}{d\lambda} \log \Gamma(k) + \frac{d}{d\lambda}(k-1) \log x_n - \frac{d}{d\lambda} \frac{x_n}{\lambda}) =$$

$$= \sum_n (-\frac{d}{d\lambda} k \log \lambda - 0 + 0 - \frac{d}{d\lambda} \frac{x_n}{\lambda}) =$$

$$= \sum_n (-\frac{k}{\lambda} + \frac{x_n}{\lambda^2}) = 0$$

Rozwiązujemy równanie.

$$\sum_n (-\frac{k}{\lambda} + \frac{x_n}{\lambda^2}) = 0$$

$$-\frac{Nk}{\lambda} + \sum_n \frac{x_n}{\lambda^2} = 0$$

$$\frac{Nk}{\lambda} = \frac{1}{\lambda^2} \sum_n x_n$$

Obie strony równania mnożymy przez λ^2.

$$Nk\lambda = \sum_n x_n$$

$$\lambda = \frac{\sum_n x_n}{Nk}$$

2.3

Obliczamy wartość estymatora dla zadanych danych.

$$\lambda = \frac{\sum_n x_n}{Nk} =$$

Wstawiamy dane.

$$\frac{1+1+2+4}{4 \cdot 0.5} = 4$$