

Тепер розглянемо випадок випадка:

Рл. Нехай задача ML опукла і β -ладна. Тоді $\forall$ розподілу $\mathcal{D}$ на $\mathbb{Z}$ алг. SGD застосований до мінімізації $L_{\mathcal{D}}$ з параметром $\eta > 0$ за T кроків повертає рез-т $\bar{w}$ такий, що $\forall w^*$

$$E(L_{\mathcal{D}}(\bar{w})) \leq \frac{1}{1-\eta\beta} \left(L_{\mathcal{D}}(w^*) + \frac{\|w^*\|^2}{2\eta T} \right)$$

$$\Rightarrow E(L_{\mathcal{D}}(\bar{w})) = E\left(L_{\mathcal{D}}\left(\frac{1}{T} \sum_{t=1}^T w^{(t)}\right)\right) \leq \left[\begin{smallmatrix} \text{опуклість} \\ \text{ліній. E} \end{smallmatrix} \right] \leq E\left(\frac{1}{T} \sum_{t=1}^T L_{\mathcal{D}}(w^{(t)})\right) \quad \text{Q.E.D.}$$

$\forall t \quad L_{\mathcal{D}}(w^{(t)}) = E_{z \sim \mathcal{D}} \ell(w^{(t)}, z)$, де ℓ - втрача функція задачі.

Позначимо z_t - t -м зразком з $\mathbb{Z}$, що випадково (вип. з $\mathcal{D}$)

обирається на t -му кроці SGD. Оскільки загалом E береться за $S = (z_1, \dots, z_T) \sim \mathcal{D}^T$, за лінійністю $\Rightarrow E\left(\frac{1}{T} \sum_{t=1}^T \ell(w^{(t)}, z_t)\right) \leq$

$\forall t$ позначимо $f_t = \ell(\cdot, z_t)$. За умовою, ці функції опуклі і β -ладні, звідси диференційовні. Тому на t -му

Кроки SGD $v_t = \nabla f_t(w^{(t)})$. З опуклості ма lem. 1,

$$\sum_{t=1}^T (f_t(w^{(t)}) - f_t(w^*)) \leq \sum_{t=1}^T \langle w^{(t)} - w^*, v_t \rangle \leq \frac{|w^*|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T |v_t|^2 \leq$$

$$\leq \left[\begin{array}{l} f_t \text{ gn. } \beta\text{-м.} \\ \geq 0 \Rightarrow \text{самообме-} \\ \text{жена, } v_t = \nabla f_t(w^{(t)}) \end{array} \right] \leq \frac{|w^*|^2}{2\eta} + \eta \beta \sum_{t=1}^T f_t(w^{(t)}), \text{ оскільки}$$

$$\frac{1}{T} \sum_{t=1}^T f_t(w^{(t)}) \leq \frac{1}{T(1-\eta\beta)} \left(\sum_{t=1}^T f_t(w^*) + \frac{|w^*|^2}{2\eta} \right), \text{ м.ч.,}$$

$$\textcircled{\leq} \mathbb{E} \left(\frac{1}{T(1-\eta\beta)} \left(\sum_{t=1}^T f_t(w^*) + \frac{|w^*|^2}{2\eta} \right) \right) = \left[\mathbb{E}(f_t(w^*)) = \right. \\ \left. = L_{\mathcal{D}}(w^*) \text{ за } \underline{\text{деф.}} \right] = \frac{1}{1-\eta\beta} \left(L_{\mathcal{D}}(w^*) + \frac{|w^*|^2}{2\eta T} \right)$$

що й потрібно. $\triangle$

Соч. 2. Нехай задача ML опукла, β -магма і β -обмежена, а її η -йя втрам ℓ задов. мови $\ell(0, z) \in [0, 1] \forall z \in \mathbb{Z}$.

Покр $\forall \epsilon > 0$ алгоритм SGD мінімізації $L_{\mathcal{D}}$, застосований з $T \geq \frac{8\beta^2\beta}{\epsilon^2}$ ітераціями (тобто го вибірка з T ек-тів, що

i.i.d. вим. го $\mathcal{D}$) і $\eta = \frac{1}{\beta(1+\frac{2}{\epsilon})}$ для $\forall$ розн. $\mathcal{D}$ на $\mathbb{Z}$ повертає

рез-т $\bar{w}$ з властивістю

$$E(L_D(\bar{w})) \leq \min_{w \in K} L_D(w) + \epsilon$$

Зокрема, ця задача еквівалентно PAC навчана.

⇒ Оск. $1 - \beta^2 = 1 - \frac{1}{1 + \frac{2}{\epsilon}} = \frac{1}{1 + \frac{\epsilon}{2}}$, за помил. P_H ,

$$E(L_D(\bar{w})) \leq \left(1 + \frac{\epsilon}{2}\right) \left(\min_{w^* \in K} L_D(w^*) + \frac{B^2}{2\eta T} \right) \leq \min_{w^* \in K} L_D(w^*) + \frac{\epsilon}{2} L_D(0) + \left(1 + \frac{\epsilon}{2}\right) \frac{B^2}{\frac{16B^2}{(1 + \frac{2}{\epsilon})\epsilon^2}} \leq \left[L_D(0) \leq 1 \text{ за умовою, } \epsilon \leq 1 \Rightarrow \frac{\epsilon}{8} \left(1 + \frac{\epsilon}{2}\right)^2 \leq \frac{\epsilon}{2} \right] \leq \min_{w^* \in K} L_D(w^*) + \epsilon,$$

що є потрібно. Навчальність зводиться до помил. розділи
(Тм. 1, і 2.) $\triangle$

Реш. Тут знову не отримали на $T=m$ ознаку того ж
характера, що для RLM, але кращу (так була $\geq \frac{96 B^2 \beta}{\epsilon^2}$).

SGD для RLM

що розб'яжються за год.

Згідно, що в задачі, RLM-алгоритмів, треба

мінімізувати $L_S(w) + \frac{\lambda}{2} |w|^2$. Якщо задача опукла, то $\forall$ навч. вибірки S L_S опукла, а отже ця сума λ -супер (сильно) опукла за ознак з Рн. попер. розділу.

Для таких ф-цій існує варіант SGD з кращою швидкістю збігності (за тж, що у Мінгій Тн.). Щоб її отримати, спочатку доведемо варіант ознак з твердженням заданого Рн. для (вз. канури) негидеренційовних ф-цій:

Лем. 4. Якщо f λ -с.оп., то $\forall u, w$ і $v \in \partial f(u)$

$$f(w) - f(u) \geq \frac{\lambda}{2} |w - u|^2 + \langle w - u, v \rangle$$

► Діємо як у дов. н.з. Рн. попер. розділу. З дов. λ -с.оп., $\forall \alpha \in (0, 1)$

$$f(w) - f(u) - \frac{\lambda}{2} (1 - \alpha) \geq \frac{\frac{|w - u|^2}{2}}{\alpha} \geq [\text{дов. суднагієнна}] \geq \frac{1}{2} \langle \alpha w + (1 - \alpha)u - u, v \rangle = \langle w - u, v \rangle. \text{ При } \alpha \rightarrow 0 \text{ отр. потрібне } \Delta$$

Для λ -с.оп. f реализуемо наступний варіант SGD з кроком проєкції на опуклому K і зі змінним кроком η_t :

$w^{(1)} := 0$, $\forall t$ обираємо v_t так, що $E(v_t | w^{(t)}) \in \partial f(w^{(t)})$,
 $w^{(t+\frac{1}{2})} := w^{(t)} - \eta_t v_t$, де $\eta_t = \frac{1}{\lambda t}$, $w^{(t+1)} := \argmin_{w \in K} |w - w^{(t+\frac{1}{2})}|$,
 результатом $\bar{w} := \frac{1}{T} \sum_{t=1}^T w^{(t)}$.

Тл. Нехай w^* - точка мінімуму λ -с.оп. ф-ції $f: K \rightarrow \mathbb{R}$, де $K \subset \mathbb{R}^d$ опукла. Виконаємо T кроків описаної вище модифікації алгоритма SGD, де припускаємо, що $E(|v_t|^2) \leq \rho^2$.

Тоді
$$0 \leq E(f(\bar{w})) - f(w^*) \leq \frac{\rho^2}{2\lambda T} (1 + \ln T)$$

► Як на початку дов. попер. Тл., з лінійності ма опуклості,

$$E(f(\bar{w})) - f(w^*) \leq E\left(\frac{1}{T} \sum_{t=1}^T (f(w^{(t)}) - f(w^*))\right) \leq$$

$\forall t$ позначимо $\nabla^{(t)} := E(v_t | w^{(t)}) \in \partial f(w^{(t)})$. За Лем. 4,

$$E(f(w^{(t)}) - f(w^*)) \leq E\langle w^{(t)} - w^*, \nabla^{(t)} \rangle - \frac{\lambda}{2} |w^{(t)} - w^*|^2 \leq$$

При цьому

$$|w^{(t)} - w^*|^2 - |w^{(t+1)} - w^*|^2 \geq \left[\begin{array}{l} w^{(t+1)} - \text{пр. } w^{(t+\frac{1}{2})} \\ \text{на } \mathcal{H}, w^* \in \mathcal{H}, \\ \text{Lem. 3} \end{array} \right] \geq |w^{(t)} - w^*|^2 - |w^{(t+\frac{1}{2})} - w^*|^2$$

("різнця квадратів")

$$-w^*|^2 \stackrel{d}{=} \langle w^{(t)} - w^{(t+\frac{1}{2})}, w^{(t)} + w^{(t+\frac{1}{2})} - 2w^* \rangle = \langle \eta_t v_t, 2(w^{(t)} - w^*) - \eta_t v_t \rangle = 2\eta_t \langle v_t, w^{(t)} - w^* \rangle - \eta_t^2 |v_t|^2$$

$$2(w^{(t)} - w^*) - \eta_t v_t \rangle = 2\eta_t \langle v_t, w^{(t)} - w^* \rangle - \eta_t^2 |v_t|^2$$

Переїдемо у цей нер-ли до очікування:

$$\textcircled{\leq} \frac{1}{2\eta_t} \mathbb{E} \left(|w^{(t)} - w^*|^2 - |w^{(t+1)} - w^*|^2 + \eta_t^2 |v_t|^2 \right) - \frac{\lambda}{2} \mathbb{E} (|w^{(t)} - w^*|^2) \leq$$

$$\leq [\text{зроба}] \leq \frac{\lambda}{2} \mathbb{E} \left((t-1) |w^{(t)} - w^*|^2 - t |w^{(t+1)} - w^*|^2 \right) + \frac{\mathcal{P}^2}{2\lambda t}$$

Накнемо, згадаємо за t , наємо за мінімізацією

$$\textcircled{\leq} \frac{1}{T} \left(\frac{\lambda}{2} \mathbb{E} \left(-T |w^{(T+1)} - w^*|^2 \right) + \frac{\mathcal{P}^2}{2\lambda} \sum_{t=1}^T \frac{1}{t} \right) \leq \frac{\mathcal{P}^2}{2\lambda T} (1 + \ln T). \triangle$$

до

Щоб застосувати цей алгоритм до мінімізації

$$f(w) = L_S(w) + \frac{\lambda}{2} |w|^2, \text{ треба ще вказати спосіб}$$

визбору вектора, оцінювання якого $\in \partial f(w^{(t)})$. Для цього дост. для випадково обраного $z \in S$ взяти $v_t \in \partial \ell(w^{(t)}, z)$, тоді

$$E(v_t + \lambda w^{(t)} | w_t) = E(v_t | w_t) + \lambda w^t \in [E(v_t | w_t) \in \partial L_S(w_t), \text{ а к-но го } L_D \text{ виме}] \in \partial f(w^{(t)}).$$

Будемо припускати $\mathcal{H} = \mathbb{R}^d$, тоді крок протекції можна не згінусовувати. $\forall t$ маємо

$$\begin{aligned} w^{(t+1)} &= w^{(t)} - \frac{1}{\lambda t} (\lambda w^{(t)} + v_t) = \frac{t-1}{t} w^{(t)} - \frac{1}{\lambda t} v_t = [a_t - a_0] = \\ &= \frac{t-1}{t} \left(\frac{t-2}{t-1} w^{(t-1)} - \frac{1}{\lambda(t-1)} v_{t-1} \right) - \frac{1}{\lambda t} v_t = \dots = [w^{(1)} = 0] = -\frac{1}{\lambda t} \sum_{i=1}^t v_i. \end{aligned}$$

Якщо задана ρ -ліній., $\forall i \ |v_i| \leq \rho \Rightarrow |w^{(t)}| \leq \frac{\rho}{\lambda} \ \forall t$,

отже $|\lambda w^{(t)} + v_t| \leq 2\rho$. З непер. Тн.:

Сок. В умовах Тн. після T кроків SGD для $f = L_S + \frac{\lambda}{2} |\cdot|^2$ (в оцінці ρ -ліній. задані)

$$0 \leq E(f(\bar{w})) - f(w^*) \leq \frac{2\rho^2}{\lambda T} (1 + \ln T).$$

Машини опорних векторів

Машини (методи) опорних векторів (support vector machines, SVM) - це клас алгоритмів лінійної бінарної класифікації точок підпростору $\mathbb{R}^d$, що максимізує відстань від розрізюючої гіперплощини до найближчих точок. Для ефективної роботи ці алгоритми вимагають умов на розподіл, які, як виявляється, дозволяють дати оцінку складності вибірки, що не залежить від d , на відміну від оцінок, що вих. VC-вимірність $d+1$ (d у опор. випадку) класу наївностей. Це важливо, якщо d велике.

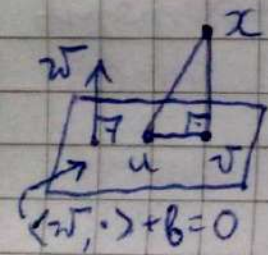
Kernel - SVM

Згадаймо, що в задачі лін. бінар. клас. ми шукаємо

дані $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ мають інтерполяцію $\{\langle w, x \rangle + b = 0\}$,
 де $w \in \mathbb{R}^d$ (вектор нормалі) і $b \in \mathbb{R}$, що $\langle w, x \rangle + b > 0$
 при $y = 1$ і $\langle w, x \rangle + b < 0$ при $y = -1$ (чим вважаємо,
 що точки дані не лежать на інтервалі.) $\Leftrightarrow y(\langle w, x \rangle + b) > 0$.
 Ця роздільність (зокр., на навч. вибірці $S = \{(x_i, y_i)\}_{i=1}^m$)
 означає, що така інтервал. існує.

Лем. Якщо $|w| = 1$, то відстань від $x \in \mathbb{R}^d$ до інтервал. з
 рівнянням $\langle w, x \rangle + b = 0$ дорівнює $|\langle w, x \rangle + b|$.

За означенням, відстань — це $\min \{ |x - v| \mid \langle v, w \rangle + b = 0 \}$.



Для $v := x - (\langle w, x \rangle + b)w$ (ортогональної проєкції
 x на інтервал.) вона дор. $|(\langle w, x \rangle + b)w| = [|w| = 1] = |\langle w, x \rangle + b|$,
 при цьому $v \in$ інтервал!

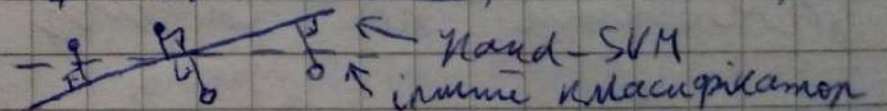
$$\langle w, v \rangle + b = \langle w, x \rangle - (\langle w, x \rangle + b) \overset{+1}{|w|^2} + b = 0.$$

Тому відстань $\leq |x-v| = |\langle w, x \rangle + b|$. З іншого боку, $\forall u$ з
 гіперпл. (тобто $\langle w, u \rangle + b = 0$):

$$|x-u|^2 = |x-v+v-u|^2 = |x-v|^2 + 2\langle x-v, v-u \rangle + |v-u|^2 =$$

$$= \left[\begin{array}{l} \langle v, w \rangle + b = 0 \\ \langle u, w \rangle + b = 0 \\ \Rightarrow \langle v-u, w \rangle = 0 \\ x-v = (\langle w, x \rangle + b)w \end{array} \right] \Rightarrow \left[\begin{array}{l} \uparrow \text{"теорема Піфагора"} \\ |x-v|^2 + |v-u|^2 \geq |x-v|^2, \text{ що й потрібно. } \triangle \end{array} \right]$$

def. SVM із жорстким розділенням (hard-SVM) - це
 алгоритм, що за розділеними даними $S = \{(x_i, y_i)\}_{i=1}^m$
 повертає гіперплоскість $\{x | \langle w, x \rangle + b = 0\}$, що правильно
 класифікує ці дані і для якої відстань від гіперплоскості
 до найближчої точки даних (т.зв. розділення, margin)
 є найбільшою можливою:



$$(w, b) \in \underset{|w|=1}{\operatorname{argmax}} \left\{ \min_{i=1, m} |\langle w, x_i \rangle + b| \mid \forall i=1, m \ y_i (\langle w, x_i \rangle + b) > 0 \right\}.$$

(Тут умова $|w|=1$ випливає з лем., формулу з якої ми виходимо)

Реш. Задача з означення максимізації на площині

$$(w, b) \in \arg \max_{|w|=1, i=\overline{1, m}} y_i (\langle w, x_i \rangle + b) \quad (*)$$

Діємо, якщо (w, b) невірно класифікує дані, то $\exists i$:

$y_i (\langle w, x_i \rangle + b) < 0$, а отже $\min_{i=\overline{1, m}} \dots < 0$. З цього розг.

вистібає, що $\exists (w, b)$, для яких це > 0 , тому максим.

не виходить на $\arg \max$. Якщо $\forall$ дана (w, b) правильно роз-
гінкує дані, то $\forall i: 0 < y_i (\langle w, x_i \rangle + b) = |y_i (\langle w, x_i \rangle + b)| = [y_i = \pm 1] \cdot$
 $= |\langle w, x_i \rangle + b|$, і отримujemy той самий вираз.

Висновком, що задачу максимізації розділення можна
замінити на задачу мінімізації $|w|$:

Реш. Ухвалі для даних $S = \{(x_i, y_i)\}_{i=1}^m$

$$(w_0, b_0) \in \arg \min \{ |w|^2 \mid \forall i = \overline{1, m} \ y_i (\langle w, x_i \rangle + b) \geq 1 \}.$$

Потім $(\frac{w_0}{|w_0|}, \frac{b_0}{|w_0|})$ загав. (*).

► Нехай (w^*, b^*) — якийсь розв'язок (*). Позн. $\gamma^* := \min_i \gamma_i$.
($\langle w^*, x_i \rangle + b^*$). Також $\forall i$

$$\gamma_i (\langle w^*, x_i \rangle + b^*) \geq \gamma^* \Rightarrow \gamma_i \left(\langle \frac{w^*}{\gamma^*}, x_i \rangle + \frac{b^*}{\gamma^*} \right) \geq 1$$

П.ч. $\left(\frac{w^*}{\gamma^*}, \frac{b^*}{\gamma^*} \right)$ задов. умовам задачі з умови Рч., отже

$$|w_0| \leq \left| \frac{w^*}{\gamma^*} \right| = [|w^*| = 1] = \frac{1}{\gamma^*}. \text{ Також } \forall i$$

$$\gamma_i \left(\langle \frac{w_0}{|w_0|}, x_i \rangle + \frac{b_0}{|w_0|} \right) = \frac{1}{|w_0|} \gamma_i (\langle w_0, x_i \rangle + b_0) \geq \frac{1}{|w_0|} \geq \gamma^*.$$

Отж. $\left| \frac{w_0}{|w_0|} \right| = 1$, це і означає нормізоване. $\triangle$

Рем. П.ч., hard-SVM може бути сформульована як
пошук найменшого $|w|^2$ за умов $\gamma_i (\langle w, x_i \rangle + b) \geq 1$.

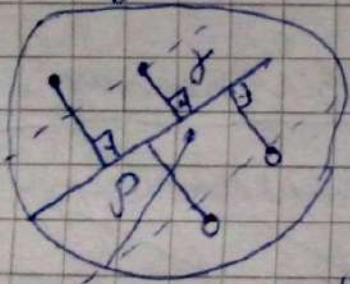
Такого роду задачі (оптимізації квадратичної ф-ції
за набору лінійних умов) звуться задачами квадратич-

ної оптимізації, що як і лінійна (LP) є окр. впадковою опуклою.

Як і раніше, тут всього можна обмежитися однокласним випадком ($b=0$), розглянувши за необхідності $(d+1)$ -випірну задачу.

Для зцінки складності вибірки знадобиться наступне:

def. Розподіл $\mathcal{D}$ на $\mathbb{R}^d \times \{\pm 1\}$ зветься розділюваним з (γ, ρ) -розділюваністю ((γ, ρ) -margin), або просто (γ, ρ) -розділюваним ((γ, ρ) -separable), якщо існує (w^*, b^*) таке, що $|w^*|=1$ і зі ім. 1 для $(x, y) \sim \mathcal{D}$ $y(\langle w^*, x \rangle + b^*) \geq \gamma^*$ і $|x| \leq \rho$.



Рем. Подто (зі ім. 1) точки знах. на $\geq \gamma^*$ відстані від класифікуючої гіперплощини та у кулі радіуса ρ . Тут знову, наприклад, геометрія

$\mathbb{R}^d \rightarrow \mathbb{R}^d : x \mapsto \lambda x$, що, ок., не впливає на задачу, зберігає $\frac{\gamma}{\beta}$.

Th. $\forall \delta \in (0, 1)$, $m \in \mathbb{N}$, $\forall$ розподілу $\mathcal{D}$ на $\mathbb{R}^d \times \{\pm 1\}$, що є (γ, δ) -розділюючим для одновимірних функцій, зі ім. $> 1 - \delta$ для $S \sim \mathcal{D}^m$ hard-SVM-алгоритм повертає w , для якого

$$L_{\mathcal{D}}(w) \leq \sqrt{\frac{4}{m} \left(\frac{\beta}{\gamma}\right)^2} + \sqrt{\frac{2}{m} \ln \frac{2}{\delta}},$$

де $L_{\mathcal{D}}$ визначає 0-1-втрачу.

► Dub. [1, Section 26.3] $\triangle$

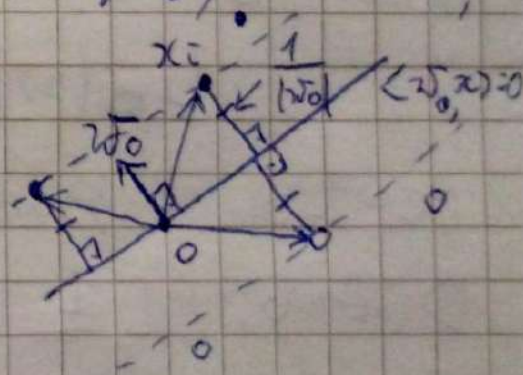
Rem. Звідси можна отримати оцінку m (складності вибір-ки), що м.ч. дійсно не залежить від d . Це можливо (і не суперечить основній теоремі), бо тут можна розподіло-обмежена умовою (γ, δ) -розділюючі, що сильніша за звичай-ну умову розділюючі (реалізовності). Разом з цим, ви-конання цієї умови часто природно припускати у практичних задачах.

Термін "опорні вектори" пояснюється наступним чином:

Тв. Нехай w_0 - розв'язок hard-SVM у одновимірній випадку:

$$w_0 \in \arg \min \{ |w|^2 \mid \forall i=1, m \quad y_i \langle w, x_i \rangle \geq 1 \}.$$

Покі w_0 є лінійною комбінацією (радісів - векторів) точок, що знаходяться на найменшій відстані від гіперплощини $\{x \mid \langle w_0, x \rangle = 0\}$ (що год. $\frac{1}{|w_0|}$ за лем.):
 $\exists \{\alpha_i\} \subset \mathbb{R}$ такі, що



$$w_0 = \sum_{\substack{i=1, m \\ y_i \langle w_0, x_i \rangle = 1}} \alpha_i x_i.$$

Такі x_i і називаються опорними векторами (support vectors).

Рем. Для розширення hard-SVM, зокр. для ядрових методів (kernel methods, див. наст. курс), буває корисно розглядати

т.зв. дуальне формулювання. Позначимо

$$g(w) := \max_{\substack{\alpha = (\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m \\ \forall i: \alpha_i \geq 0}} \sum_{i=1}^m \alpha_i (1 - y_i \langle w, x_i \rangle) = \begin{cases} 0, & \forall i: y_i \langle w, x_i \rangle \geq 1 \\ +\infty & \text{if im. van.} \end{cases}$$

Попри розб'язок оптимізації hard-SVM можна екв-но переписати як

$$\min_w \left(\frac{1}{2} |w|^2 + g(w) \right) = \min_w \max_{\alpha \geq 0} \left(\frac{1}{2} |w|^2 + \sum_i \alpha_i (1 - y_i \langle w, x_i \rangle) \right) \Leftrightarrow$$

Бо для мінімізації необхідно $g(w) = 0$, тобто виконання умов $y_i \langle w, x_i \rangle \geq 1 \forall i$, $\min$ і $\max$ можна переписати за наступним правилом:

Попр. $\forall \varphi$ -yii $f: X \times Y \rightarrow \mathbb{R}$

$$\min_{x \in X} \max_{y \in Y} f(x, y) \geq \max_{y \in Y} \min_{x \in X} f(x, y).$$

Виявляється, що у даному випадку ця нер-сть буде рівністю:

$$\Leftrightarrow \max_{\alpha \geq 0} \min_w \left(\frac{1}{2} |w|^2 + \sum_i \alpha_i (1 - y_i \langle w, x_i \rangle) \right) \Leftrightarrow$$

Тепер $\min_w$ ми можемо отримати диференціальну φ -yii

fig w , тому m мінімум буде критична, де її градієнт $= 0$: $w - \sum_i \alpha_i y_i x_i = 0$. Підставимо w звідси:

$$\Leftrightarrow \max_{\alpha \geq 0} \left(\frac{1}{2} \left| \sum_i \alpha_i y_i x_i \right|^2 + \sum_i \alpha_i \left(1 - y_i \left\langle \sum_j \alpha_j y_j x_j, x_i \right\rangle \right) \right) = \left[\begin{array}{c} \text{розкр.} \\ \text{квадрат} \end{array} \right] =$$

$$= \max_{\alpha \geq 0} \left(\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \right).$$

Це i є квадратна задача.

Soft-SVM

"М'якість" означає послаблення умов hard-SVM $y_i(\langle w, x_i \rangle + b) \geq 1$ на деякі додаткові змінні ξ_i , середнє значення яких ми хочемо мінімізувати разом з $|w|^2$:

def. SVM із м'яким розділенням (Soft-SVM) з параметром

$\lambda > 0$ - це алгоритм, що за даними $S = \{(x_i, y_i)\}_{i=1}^m$ повертає гіперплощину $\{x \mid \langle w, x \rangle + b = 0\}$, що задовольняє умову

$y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i$, $i = \overline{1, m}$ для данных $\xi_i \geq 0$, i для
 яких найменше можливе значення приймає функція

$$\lambda |w|^2 + \frac{1}{m} \sum_{i=1}^m \xi_i$$

Роз. Подібно тут мінімізуємо одночасно $|w|$ та ξ_i , а
 параметр λ регулює "компроміс" між цими задачами. Дані
 тепер можуть бути й нерозділюваними. Зауважимо, що $\forall i = \overline{1, m}$
 найменшим можливим ξ_i буде:

$$\begin{cases} 0, & y_i(\langle w, x_i \rangle + b) \geq 1 \\ 1 - y_i(\langle w, x_i \rangle + b) & \text{в інш. вар.} \end{cases} = \max \{0, 1 - y_i(\langle w, x_i \rangle + b)\} =$$

$= \ell_{\text{hinge}}((w, b), (x_i, y_i))$ (наш розглядати одновимірний варі-
 ант цієї ϕ - y_i функції), отже $\frac{1}{m} \sum_{i=1}^m \xi_i = \frac{1}{m} \sum_{i=1}^m \ell_{\text{hinge}}((w, b), (x_i, y_i)) = \frac{1}{m} \sum_{i=1}^m \ell_{\text{hinge}}((w, b), (x_i, y_i)) = L_S(w, b)$.

Сол. Для даних $S = \{(x_i, y_i)\}_{i=1}^m$ будь-який розв'язок задачі
 $(w_0, b_0) \in \text{argmin} (\lambda |w|^2 + L_S(w, b))$,

де L_S обчислюється на S відповідно до ф-ції втрач ℓ_{hinge} ,
буде розв'язком soft-SVM на S для параметра λ .

Попри отриманню RLM для опуклої ліній. задачі. Дійсно,
обмежена отриманий випадком, тобто $b=0$ (як за-
давалося, це не зменшує загальність). Раніше вже пере-
вірили, що ℓ_{hinge} , а отже L_S , опукла.

Впр. Показати, що $\ell_{\text{hinge}}(w, (x, y)) = \max\{0, 1 - y \langle w, x \rangle\}$
 $|x|$ -лінійна (за w) $\forall x, y$.

П.ч. задача буде опуклою ρ -лінійною на $X \times \{\pm 1\}$
де $X = \mathcal{D}_\rho(0) = \{x \mid |x| \leq \rho\}$; якщо шукати розв'язки
у $\mathcal{K} = \mathcal{D}_B(0) = \{w \mid |w| \leq B\}$, вона буде ще й обмеженою.
З результатів розділу про RLM безпосередньо ви-
пливає наступне (таким тут використано $\ell_{0-1} \leq \ell_{\text{hinge}}$):

Сок. Для гомогенного розподілу $\mathcal{D}$ на $\mathcal{D}_\beta(0) \times \{\pm 1\}$, $m \in \mathbb{N}$,
 результатом $A(S)$ роботи V -soft-SVM алгоритма A ^{з нах. до} на набір
 вибірки $S \sim \mathcal{D}^m$ задовольняє умови: $\forall u \in \mathbb{R}^d$

$$E(L_{\mathcal{D}}^{0-1}(A(S))) \leq E(L_{\mathcal{D}}^{\text{hinge}}(A(S))) \leq L_{\mathcal{D}}^{\text{hinge}}(u) + \lambda |u|^2 + \frac{2\beta^2}{\lambda m},$$

 де помилки $L_{\mathcal{D}}^{0-1}$, $L_{\mathcal{D}}^{\text{hinge}}$ обчислюються відносно ℓ_{0-1} , ℓ_{hinge} відносно ℓ_{hinge} .
 Якщо не указати розв'язки у $\mathcal{K} = \mathcal{D}_\beta(0)$ і покласти $\lambda = \sqrt{\frac{2\beta^2}{B^2 m}}$,

то $\forall \mathcal{D}, m, S \sim \mathcal{D}^m$

$$E(L_{\mathcal{D}}^{0-1}(A(S))) \leq E(L_{\mathcal{D}}^{\text{hinge}}(A(S))) \leq \min_{|w| \leq B} L_{\mathcal{D}}^{\text{hinge}}(w) + \beta B \sqrt{\frac{8}{m}}.$$

Реш. Остання частина тут - з Th. 1. розгляду про RLM. Як
 бачимо, оцінку на m зразку (скажімо вибірку) знову
 можна дати незалежно від d , якщо обмежити "розмір" дані:

$$m \geq \frac{8\beta^2 B^2}{\epsilon^2}$$
 для бажаної помилки ϵ .

Ек. Розглянемо задачу виявлення належності тексту до певної теми (належить / не належить) за допомогою bag-of-words підходу: текст представляється вектором у $\{0, 1\}^d$, де d - кількість усіх можливих слів (довжина словника), а координата i вказує на наявність слова у тексті. Тоді J^2 - це кількість у тексті різних слів ($\leq$ його довжина). B^2 - це кількість значущих слів, наявність яких може говорити про належність тексту до теми і які т.ч. викликають як "важ" 1 (замість 0) у $\langle w, x \rangle$. Тоді оцінка, що лінійна за $J^2 B^2$, як у поперед. Соч., може бути на порядок кращою за лінійну за d з VC-визірності (максимо, $100 \times 100 = 10\,000$ проти $100\,000$).

Реш. Як було вказано наприкінці поперед. розділу, ми

можливо застосувати SGD для описаної RLM-задачі. Як
 було там показано, на t -му кроці цього алгоритма правдо
 оновлення приймає вигляд $w^{(t+1)} = -\frac{1}{\lambda t} \sum_{j=1}^t \sigma_j$, де σ_j -
ф-ції втрат $y w^{(j)}$ на випадково обраному $(x_i, y_i) \in S$. Як
 правило бачимо, що $\sigma_j = 0$ при $y_i < w^{(j)}, x_i > 1$ і
 $-y_i x_i$ у іншому випадку. Т.ч., отримуюємо алгоритм
 (для пар. $\lambda > 0$ і кінц. кроків T), де позн. $\theta^{(t)} := -\sum_{j=1}^t \sigma_j$:

- $\theta^{(1)} := 0$

- $\forall t = 1, T$:

$w^{(t)} := \frac{1}{\lambda t} \theta^{(t)}$;

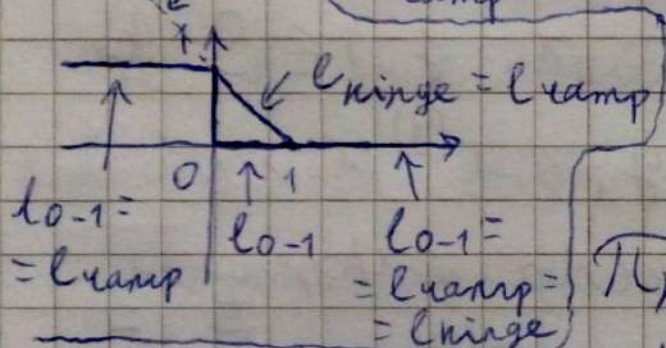
$i \sim U(m)$;

$\theta^{(t+1)} := \begin{cases} \theta^{(t)} & y_i < w^{(t)}, x_i > 1; \\ \theta^{(t)} - y_i x_i & \text{у інш. ван.}; \end{cases}$

- повертаємо $\bar{w} := \frac{1}{T} \sum_{t=1}^T w^{(t)}$.

Рем. Крим l_{0-1} і l_{hinge} мають розглядати наступну
оп-цію втрам (ramp loss):

$$l_{\text{hinge}} l_{\text{ramp}}(\tilde{w}, (x, y)) := \min \{1, l_{\text{hinge}}(\tilde{w}, (x, y))\}.$$



Для неї $l_{0-1} \leq l_{\text{ramp}} \leq l_{\text{hinge}}$, тобто
 вона є найкращим оцінювачем l_{0-1} .

При цьому, на $\text{fig. fig } L^{\text{ramp}}_{\oplus}$, для
 $\text{fig. помилки } L^{\text{hinge}}_{\oplus}$, як і для $L^{\text{hinge}}_{\oplus}$, можна оцінити
 порівняно $\sqrt{\frac{D^2 B^2}{m}}$ (бо, як $L^{\text{hinge}}_{\oplus}$ і на $\text{fig. fig } L^{\text{ramp}}_{\oplus}$, вона
 "фігурує" "масштаб" $\sqrt{D}$). Втім, оскільки вона
 неопукла (див. рис.), для неї не існує ефективних
 алгоритмів оптимізації.