

On Cost-Sensitive Support Vector Machines (aihe-esittely)

Terho Mutikainen

16.6.2023

Ohjaaja: *DI Aleksi Avela*

Valvoja: *Prof. Pauliina Ilmonen*

Työn saa tallentaa ja julkistaa Aalto-yliopiston avoimilla verkkosivuilla. Muilta osin kaikki oikeudet pidätetään.

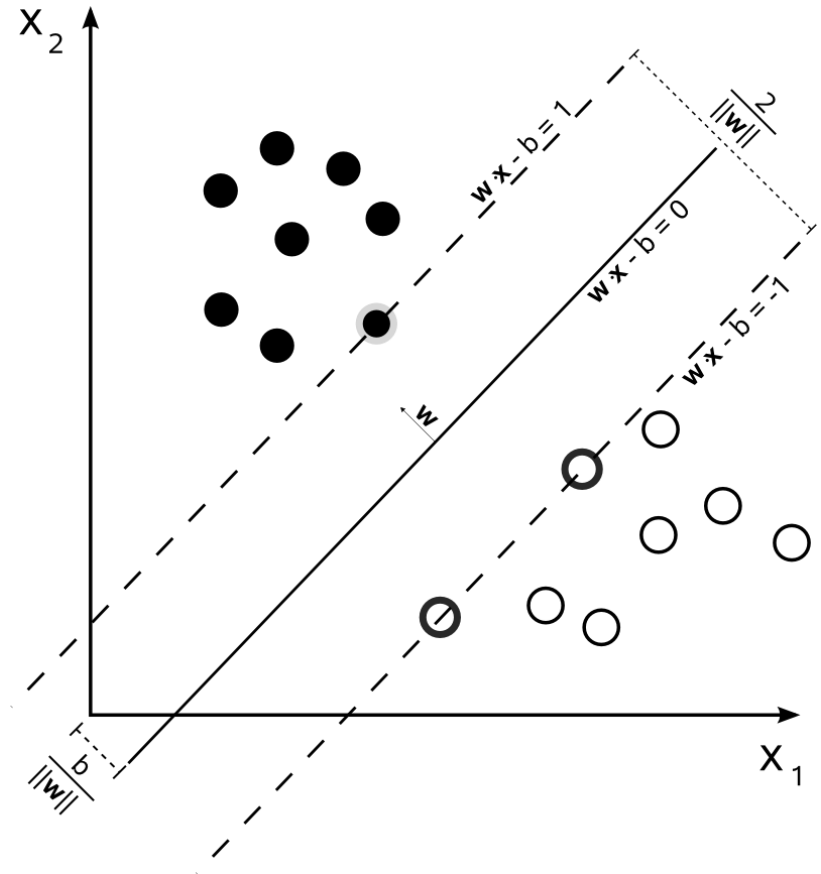
Sisällysluettelo

- Tausta
- Tavoitteet
- Rajaukset
- Aineistot & Menetelmät
- Aikataulu

TAUSTA

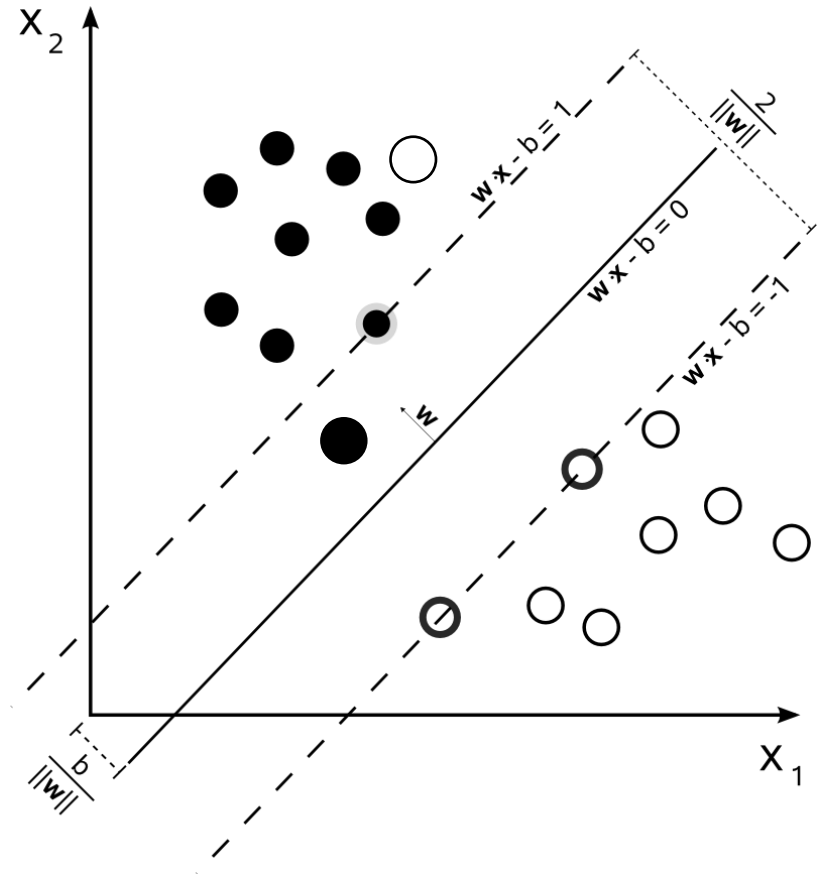
Soft-margin SVM

- Valvottu koneoppimisalgoritmi
- Sovittaa *parhaiten* luokat erottavan hypertason
- [Kuva](#): Wikimedia-käyttäjä Cyc, 2008



Soft-margin SVM

- Valvottu koneoppimisalgoritmi
- Sovittaa *parhaiten* luokat erottavan hypertason
- [Kuva](#): Wikimedia-käyttäjä Cyc, 2008



Epätasaisen datan haasteet

- Epätasaisuus: $N(+) \ll N(-)$
- SVM on puolueellinen enemmistöluokkaa (–) kohtaan
- Vähemmistöluokan (+) instanssit luokitellaan heikommin
- Ongelma, jos vääristä negatiivisista syntyy suuri *kustannus*, $C_{FN} > C_{FP}$

Kustannusmatriisi	Todellinen –	Todellinen +
Ennustettu –	0	C_{FN}
Ennustettu +	C_{FP}	0

Kustannukset huomioiva oppiminen

- *Jos kustannukset tiedetään*, halutaan luokittelijan minimoivan ne valitsemalla luokka i havainnolle $\mathbf{x}$.
- $\min_i \mathbb{E}[C \mid \mathbf{x}] = \min_i (\mathbb{P}[FN \mid \mathbf{x}]C_{FN} + \mathbb{P}[FP \mid \mathbf{x}]C_{FP})$



TAVOITTEET

Tavoitteet

- Tarkastella epätasaisen datan aiheuttamia ongelmia
- Tutkia mahdollisuuksia SVM:n parantamisesta väärinluokittelukustannuksilla
 - Algoritmin muokkaus
 - Luokitteluaineiston muokkaus
- Pohtia muokkausten ja kustannusten tarkoituksenmukaisuutta

RAJAUKSET

Rajaukset

- Binäärinen luokittelu
- Kustannusmenetelmät
- Sisäsyntyiset erot luokkien yleisyyden välillä
- SVM:n lineaarinen kerneli

AINEISTOT JA MENETELMÄT

Aineistot

- Simuloitu data
- [kaggle.com](https://www.kaggle.com)

Menetelmät

- Python-ohjelmointikieli
 - Sklearn
 - Imblearn
- Mahdollisesti Triton-laskentaklusteri

AIKATAULU

Aikataulu

Kesäkuu

- Aiheen esittely
- Aiheeseen perehtyminen

Heinäkuu

- Aiheeseen perehtyminen
- Simulointien tekeminen
- Kirjoittaminen, 1. version palautus

Elokuu

- Viimeistely ja työn esittely

VIITTEET

Viitteet

1. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9), 1263-1284.
2. Elkan, C. (2001). The foundations of cost-sensitive learning. In International joint conference on artificial intelligence (Vol. 17, No. 1, pp. 973-978).
3. Weiss, G. M., McCarthy, K., & Zabar, B. (2007). Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?. Dmin, 7(35-41), 24.