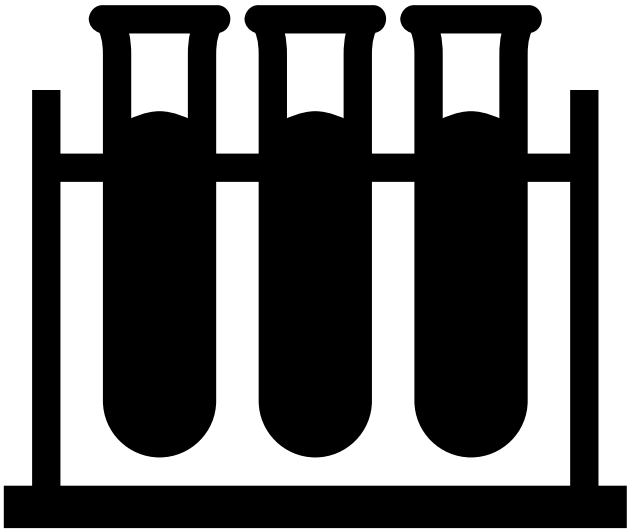


Postavka eksperimenta

Tehnike evaluacije

- Glavno načelo procene kvaliteta modela:
  - Podaci korišćeni za procenu kvaliteta modela ni na koji način ne smeju biti upotrebljeni prilikom treninga
  - Deluje jednostavno, ali se u praksi ispostavlja kao vrlo pipavo
- Još jedno načelo procene kvaliteta modela:
  - Trening i test skup treba da imaju istu raspodelu kao i buduća opažanja
  - Deluje pipavo i jeste pipavo

Tehnike evaluacije

- Podela na trening i test skup
  - Sve je u redu ukoliko je skup podataka vrlo velik i reprezentativan, ali u suprotnom…
  - Kako izvršiti podelu?
  - Šta ako su raspodele trening i test skupa različite?
  - Velika varijansa ocene greške

- Unakrsna validacija
  - Računski zahtevna
  - Kako odabrati K?
  - Sve instance se koriste u proceni kvaliteta, pa je pouzdanija, ali i dalje jedan sloj ne mora imati istu raspodelu kao preostali

Tehnike evaluacije

- Kako ublažiti ovaj problem?
- Koristiti velike količine podataka
- Koristiti napredne tehnike uzorkovanja
- Stratifikacija

Stratifikacija

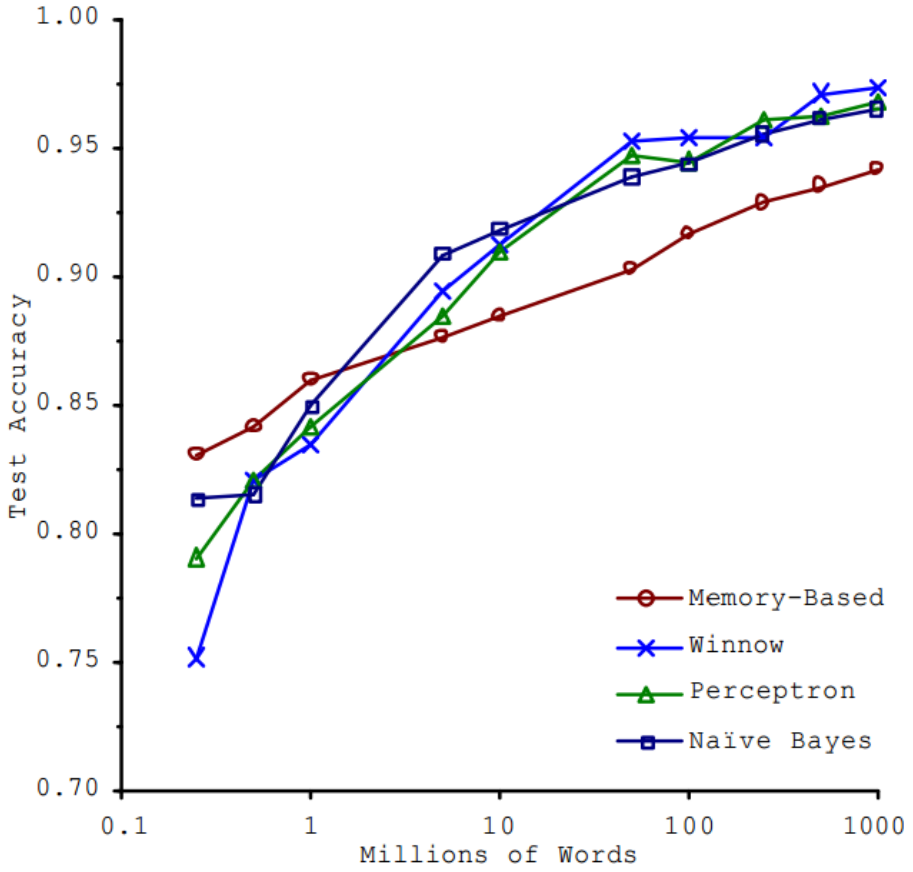
- Prilikom deljenja podataka, obezbediti da delovi imaju istu raspodelu kao i ceo skup podataka
- Teško za male skupove podataka
- Pojednostavljena varijanta: očuvati raspodelu ciljne promenljive
  - Sortirati podatke u odnosu na ciljnu promenljivu
  - Ako je *K* broj delova, neka instance sa indeksima $i + j \cdot K$ čine deo P_i za $i = 1, \ldots, K$ i $j = 0,1, \ldots$

x_1	x_2	x_3	y
0	6	2	1
7	2	3	4
1	3	1	5
6	5	1	5
4	9	7	6
3	3	4	6
1	1	6	7
7	2	1	7
1	9	0	8
2	9	9	9

Na koliko podataka da treniramo model?

Banko, M. and Brill, “Scaling to very very large corpora for natural language disambiguation”, 2001.

- Zadatak: klasifikacija reči koje se često zamenjuju
  - npr. {to, two, too}, {then, than}
  - For breakfast I ate _____ eggs.
- Proučavali su efekat korišćenja različitih obučavajućih algoritama na različitim veličinama skupa podataka



“It’s not who has the best algorithm that wins. It’s who has the most data”

Obrazloženje velike količine podataka

- Pretpostavimo da imamo veliku količinu podataka
- Pretpostavimo da obeležja $x \in \mathbb{R}^{n+1}$ sadrže dovoljno informacija da može tačno predvideti *y*
  - Primer: u rečenici *For breakfast I ate* ____ *eggs* nam okolnje reči daju dovoljno konteksta da odlučimo da je nedostajuća reč *two* a ne *too* ili *to*
- Pretpostavimo da koristimo obučavajući algoritam sa mnogo parametara
  - Npr. logistička ili linearna regresija sa mnogo obeležja
  - Ovakvi algoritmi imaju malo sistematsko odstupanje

    $\rightarrow J_{train}(\theta)$ će biti malo
  - Ako imamo veliki trening skup, malo je verovatno da će ovakav algoritam biti overfitovan

    $\rightarrow J_{train}(\theta) \approx J_{test}(\theta)$
  - Dakle, verovatno je da je $J_{test}(\theta)$ malo

Finalni saveti

- Proučiti postojeće algoritme
- Ustanoviti zašto ne daju dobre rezultate
- Proveriti da li je forma modela adekvatna
- Proveriti da li je funkcija greške adekvatna
- Proveriti da li se regularizacija može izmeniti kako bi nametnula adekvatnu strukturu modela
- Proveriti da li se optimizacioni metod može zameniti bržim
- Proveriti da li se optimizacioni problem moze aproksimirati
- Pravilno uraditi selekciju i evaluaciju modela
- Analizirati dobijeni model