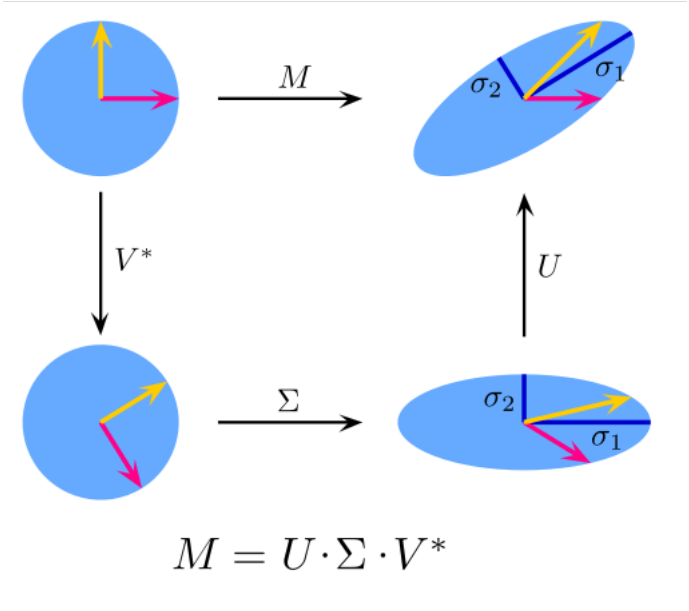
Singularna dekompozicija matrice

Singularnom dekompozicijom matrice (SVD dekompozicijom) napravljena je razdioba rečenica na kategorije. Svaki je komentar razdvojen na rečenice te je tada svakoj rečenici određeno opisuje li okus, miris, izgled ili osjećaj.
Skup za učenje SVD klasifikatora definiran je u prethodnom poglavlju. Za ovu je klasifikaciju izabrana SVD dekompozicija zbog vrlo velike brzine u odnosu na bilokoju metodu strojnog učenja. U nastavku slijedi objašnjenje singularne dekompozicije matrica i njene primjene na dani problem.
Općenito, SVD matrice je rastav matrice $X, X \in M_{m \times n}$ na:

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

gdje su U i V $m \times m$ i $n \times n$ ortonormirane matrice, a Σ je $m \times n$ dijagonalna matrica s nenegativnim singularnim vrijednostima σ_j, $j = 1, \cdots min(m, n)$, koje se najčešće zapisuju sortirane padajući po dijagonali. Slika 1 nudi objašnjenje SVD-a pomoću grafičkog prikaza. S jedne strane, objašnjenno je kako matrica M djeluje na bazne vektore nekog prostora i transformira kružnicu u zarotiranu elipsu. S druge strane tu matricu moguće je razdvojiti na tri matrice gdje su prva i zadnja ortonormirane matrice koje djeluju kao matrice rotacija, dok srednja matrica skalira vektore baze na različite vrijednosti, odnosno "rasteže" kružnicu u elipsu. Iz toga se lako vidi da su singularne vrijednosti matrice upravo duljine poluosi novonastale elipse.



Slika 1: Singularna dekompozicija matrice

Singularnu dekompozicij matrice U tu su svrhu određene značajke na način opisan u prethodnim poglavljima. Pretpostavljamo da je rečenica u kontekstu pripadne kategorije reprezentirana kao niz kojeg opisuju izabrane značajke, f_j.

Na taj smo način rečenice smjestili u vektorski prostor veličine m, gdje je m broj značajki. Tada, koristeći SVD, želimo naći potprostor kojeg zauzima svih n rečenica unutar skupa za učenje. Postupak gradnje klasifikatora ponovljen je za sve četiri kategorije zasebno. Svaki od tih postupaka vratio je potprostor koji određuje kategoriju i singularne vrijednosti koje određuju koliko se dokumenti u kategoriji razlikuju po baznim vektorima (poluosi elipse koja određuje upada li rečenica u kategoriju). Jednostavnosti radi, daljnji postupak bit će opisan samo na jednoj kategoriji, npr. okusu.

Unutar kategorije okusa definiramo prosječnu kategoriju koja sadrži glavnu informaciju za rečenice koje opisuju okus:

$$a = \frac{1}{n} \sum_{j=1}^{n} f_j, \tag{1}$$

Sada je devijacija rečenice od prosječne jednaka:

$$x_j = f_j - a. \tag{2}$$

Sada vidimo da će vektorski prostor sa središtem u točki a puno točnije opisivati traženi prostor. Naime, bit će osjetljiviji na male razlike, jer ako je a daleko od središta koordinatnog sustava, svi vektori koji su u približno jednakom smjeru, činit će se kao da su vrlo blizu. Baza prostora razapetog stupcima matrice X sastoji se od stupaca matrice U koji odgovaraju singularnim vrijednostima različitim od nule. Naime, singularne vrijednosti jednake nuli znače da se rečenice iz te kategorije u pripadajućem smjeru ne razlikuju od nula, odnosno da je taj smjer okomit na potprostor koji ih razapinje. U ovom je algoritmu korištena vrijednost 10^{-3} kao granica ispod koje smatramo da su singularne vrijednosti dovoljno da bi utjecaj pripadnih vektora bio zanemariv.

Sažeto, klasifikator je istreniran na sljedeći način:

- Konstruirana je matrica X u kojoj je svaki stupac prikaz jedne rečenice pomoću značajki
- Matrica je normalizirana
- Napravljena je SVD dekompozicija nad matricom
- Matrici U je "odrezano" zadnjih k vektora, s obirom na pripradne singularne vrijednosti

Sada za proizvoljnu rečenicu f, radimo projekciju vektora $f - a$ na prostor razapet stupcima matrice $U_\sigma = [u_1, \ldots, u_\sigma]$. Drugim riječima, želimo riješiti $U_\sigma y = f - a$, iz čega slijedi

$$y = U_V^T (f - a).$$

Ponavljanjem postupka za svaku kategoriju, dobivamo četiri klasifikatora i radimo projekciju rečenice na svaki od tih prostora. Tada gledamo udaljenosti od središta elipsa koje opisuju kategorije. Naravno, te je udaljenosti potrebno gledati normirano s obzirom na singularne vrijednosti(veće singularne vrijednosti znače da se rečenice koje opisuju tu kategoriju dosta razlikuju u smjeru pripadnog vektora, pa moramo dopustiti i novoj rečenici da po toj vrijednosti

dosta odstupa). Dakle, promatramo $\sum(y_i - \sigma_i)^2$.

Tablica 1: Matrica konfuzije za klasifikaciju po kategorijama

	Izgled	**Okus**	**Miris**	**Osjećaj**
Izgled	806	169	15	10
Okus	2	922	65	11
Miris	13	594	384	9
Osjećaj	49	361	11	579

U tablici 1 prikazana je matrica konfuzije dobivena pokretanjem algoritma na nasumično odabranih 1000 rečenica.

To je bio prvi korak pri određivanju kategorije kojoj rečenica pripada. Ipak, iako je izračunata prikazna matrica konfuzije, daljnjom analizom zaključili smo da rečenice često istovremeno opisuju više od jedne kategorije, pa je ponekad potrebno dopustiti da jedna rečenica opisuje više kategorija. Stoga smo definirali funkciju dobrote:

$$f(s) = |TP| - \frac{1}{3}|FP|,$$

gdje TP označava True pozitive rečenice, a FP False positive rečenice. Maksimiziranjem te funkcije dobivene su optimalne dopuštene udaljenosti od središta elipse za svaku kategoriju.

Implementacija

Konstruktor klase Svd može primiti značajke i listu komentara za treniranje ili adresu lokacije s koje treba učitati matrice U i Σ. U prvom slučaju, konstruktor poziva funkciju **returnMatrix** koja vraća matricu X i zatim prolazi po gore opisanom algoritmu i računa SVD. Klasa sadrži još i metodu **reduceUandS** koja reducira dimenzije matrice U i Σ, po već objašnjenom kriteriju, metodu **projection** koja prima rečenicu i radi projekciju na potprostor klasifikatora, te funkciju **save me** koja u datoteku sprema izračunati klasifikator, odnosno točnije, matrice U i Σ.