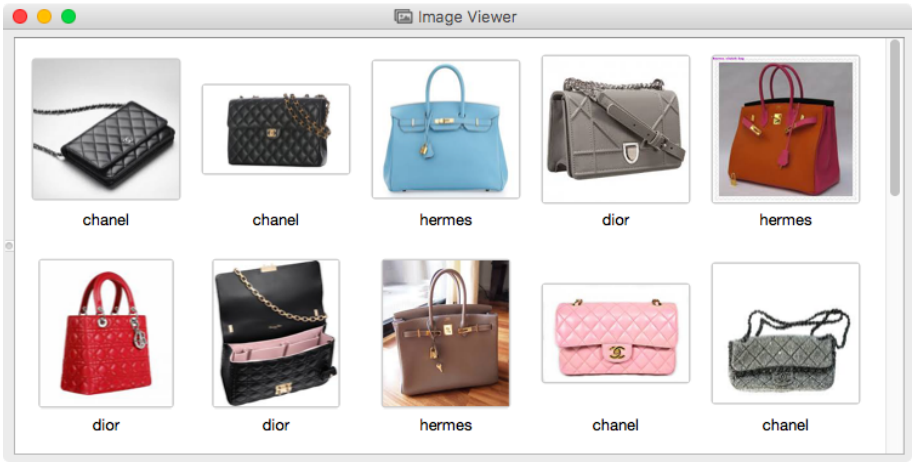
Poglavje 3

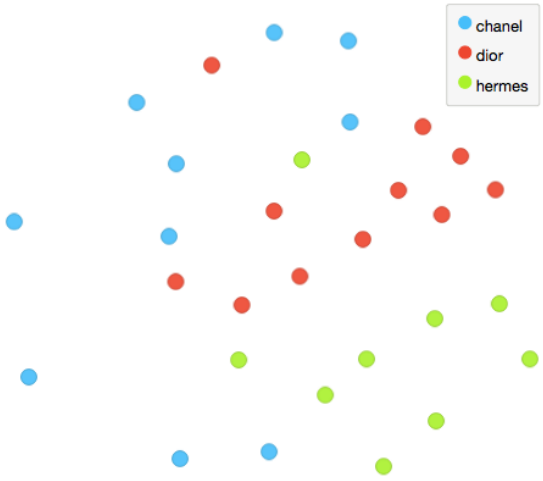
Projekcije in zmanjšanje dimenzionalnosti podatkov

Modeli, ki jih gradimo v strojnem učenju, povzemajo podatke tako, da v nekem formalnem zapisu predstavijo glavne vzorce, ki so te podatke oblikovali. V primeru večdimenzionalnih podatkov nas tako na primer zanima, ali bi podatke lahko popisali z manjšim številom spremenljivk. Z izborom značilk smo se ukvarjali že v prejšnjem poglavju, a nas bo tu zanimalo, če lahko podatke predstavimo z novimi značilkami, ki bi podatke lahko predstavili bolj kompaktno, pri tem odkrili kakšno zanimivo preslikavo starih v nove značilke, ter morda celo dosegli, da je novih značilk izjemno malo. Recimo, dve, tako da lahko vse primere izrišemo v razsevnem diagramu in tam s pomočjo vizualizacije razmišljamo o njihovih podobnosti in strukturi prostora primerov. Ljudje smo vizualna bitja in nam izris kart primerov lahko intuitivno pove več kot pa recimo matematične enačbe, ki morda te povezujejo.

Začnimo s primerom. Na sliki 3.1 so ženske torbice različnih proizvajalcev. Z uporabo že zgrajenih globokih mrež za razvrščanje slik lahko slike torbic pretvorimo v vektorje. Mreža Inception v3 nam na svojem predzadnjem nivoju na primer za vsako sliko vrne vektor dolžine 2048, oziroma sliko popiše z 2048 atributi. S postopki, ki jih bomo opisali v tem poglavju, lahko take opise zmanjšamo in predstavimo torbice v nizko dimenzionalnem prostoru. Slika 3.2 tako prikazuje predstavitev primerov v dvodimenzionalnem prostoru, kjer je razvidno, da ena Hermesova torbica bolj podobna torbicam ostalih proizvajalcev, in da se je Dior pri eni od torbic zgledoval po torbicah Chanela. Morda. Vsekakor bi ta opažanja morali preveriti, a je zanimivo, kako hitro nas vizualizacije navdahnejo z idejami. Prav zato so vizualizacije v odkrivanju znanj iz podatkov tako pomembne in zato so pomembne tudi tehnike zmanjšanja dimenzionalnosti in odkrivanja nizko dimenzionalnih projekcij podatkov.



Slika 3.1: Ženske torbice različnih proizvajalcev.



Slika 3.2: Predstavitev ženskih torbic s slike 3.1 v dvodimenzionalnem prostoru.

3.1 Metoda glavnih komponent

Prva, morda tudi najbolj znana metoda za zmanjšanje dimenzij in odkrivanje zanimivih projekcij podatkov je metoda glavnih komponent. Predpostavimo, da imamo dano matriko učnih primerov $X \in \mathbb{R}^{m \times n}$, kjer je m število primerov in n število atributov. Primer $x^{(i)} \in X$, torej i-ti primer v učni množici primerov, bo tako opisan z n atributi $x \in \mathbb{R}^n$, za katere bomo tu privzeli, da so njihove vrednosti realna števila. V splošnem iščemo projekcije podatkov tako, da atributni zapis primerov z n atributi nadomestimo z atributnim zapisom z k novimi atributi tako, da je $k \ll n$ in da nam novi atributi povedo čim več o primerih iz učne množice.

Pričnimo s $k = 1$, torej s projekcijo v eno samo dimenzijo. Dogovorimo se, da bo naša projekcija linearna in da bomo vrednost novega atributa tvorili kot linearno kombinacijo originalnih atributov. Smer projekcijske ravnine oziroma smer premice, na katero bomo projicirali podatke označimo z enotskim vektorjem u_1, za katerega torej velja $u_1^\intercal u_1 = 1$.

Naj bo $x^{(i)}$ primer iz učne množice. Njegova projekcija na projekcijsko ravnino je skalar, $u_1 x^{(i)} \in \mathbb{R}$. Naj bo $\bar{x}$ središčna točka podatkov:

$$\overline{x} = \frac{1}{m} \sum_i x^{(i)}$$

Projekcija središčne točke na projekcijsko ravnino je $u_1 \overline{x}$. Primeri v učni množici odstopajo od središčne točke, eni bolj, drugi manj. Primere bi želeli projicirati na ravnino tako, da ta odstopanja čim bolj zajamemo, torej, da so tudi v projicirani ravnini čim večja. Povprečno kvadratno odstopanje v projekciji od projekcije središčne točke imenujemo tudi varianca točk v projekciji:

$$\mathrm{Var}(u_1^\intercal X^\intercal) = \frac{1}{n} \sum_{i=1}^{m} \left(u_1^\intercal x^{(i)} - u_1^\intercal \overline{x} \right)^2$$

Z drugimi besedami, želimo poiskati tak projekcijski vektor u_1, ki maksimizira varianco projekcije $\mathrm{Var}(u_1^\intercal X^\intercal)$. Poigrajmo se malce z enačbo za varianco projekcije:

$$\begin{aligned} \mathrm{Var}(u_1^\intercal X^\intercal) &= \frac{1}{m} \sum_{i=1}^{m} \left(u_1^\intercal x^{(i)} - u_1^\intercal \overline{x} \right)^2 \\ &= \frac{1}{m} \sum_{i=1}^{m} \left(u_1^\intercal x^{(i)} u_1^\intercal x^{(i)} - 2 u_1^\intercal x^{(i)} u_1^\intercal \overline{x} + u_1^\intercal \overline{x} u_1^\intercal \overline{x} \right) \end{aligned} \tag{3.1}$$

Tu upoštevamo, da je $u_1^\intercal x$ skalarni produkt dveh vektorjev, in da je transponirana vrednost realnega števila enaka temu številu. Torej je na primer $u_1^\intercal x = (u_1^\intercal x)^\intercal = x^\intercal u_1$. Z upoštevanjem

tega se nam zgornji izraz primerno poenostavi:

$$\begin{aligned}\mathrm{Var}(u_1^\intercal X^\intercal) &= \frac{1}{m}\sum_{i=1}^{m} u_1^\intercal\Big(x^{(i)}x^{(i)\intercal} - 2x^{(i)}\overline{x}^\intercal + \overline{x}\,\overline{x}^\intercal\Big)u_1 \\ &= u_1^\intercal\Big(\frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \overline{x})(x^{(i)} - \overline{x})^\intercal\Big)u_1\end{aligned} \tag{3.2}$$

V oklepaju je kovariančna matrika! Prejšnji stavek smo končali s klicajem zato, ker je kovarianca znan in mnogokrat uporabljan koncept v statistiki in nam pove, kako sta dve naključni spremenljivki povezani. Označimo kovariančno matriko s S:

$$\mathrm{S} = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \overline{x})(x^{(i)} - \overline{x})^\intercal \tag{3.3}$$

V našem primeru imamo n atributov, kar pomeni, da je kovariančna matrika dimenzije $\mathrm{S} = \mathbb{R}^{n\times n}$. Zapišimo še enkrat dobljeno enačbo za varianco:

$$\mathrm{Var}(u_1^\intercal X^\intercal) = u_1^\intercal \mathrm{S} u_1 \tag{3.4}$$

Spomnimo, da je naš namen poiskati projekcijski vektor u_1 pri katerem je zgornja varianca maksimalna. Za vektor u_1 zahtevamo, da je to enotni vektor (sicer maksimizacija zgornje enačbe ne bi imela smisla, saj bi to dosegli z neskončno dolgim vektorjem u_1). Maksimizacijo variance, kjer iščemo ustrezen vektor u_1 z omejitvijo $u_1^\intercal u_1 = 1$ rešimo z uporabimo Lagrangeovih multiplikatorjev. Maksimiziramo torej izraz:

$$f(u_1) = u_1^\intercal \mathrm{S} u_1 + \lambda_1(1 - u_1^\intercal u_1) \tag{3.5}$$

V maksimumu bo odvod zgornje enačbe enak nič:

$$\frac{\partial f(u_1)}{\partial u_1} = \mathrm{S}u_1 - \lambda_1 u_1 = 0 \tag{3.6}$$

kar pomeni,

$$\mathrm{S}u_1 = \lambda_1 u_1 \tag{3.7}$$

Tu je S matrika katere vrednosti poznamo, u_1 je vektor, ki ga iščemo, λ_1 pa skalar. Poznano? Seveda! Zgornjemu pogoju ustreza le tak u_1, ki je lastni vektor matrike S. Ampak korelacijska matrika ima več lastnih vektorjev. Kateri med njimi je pravi? Množimo zgornjo enačbo z $u_1^\intercal$:

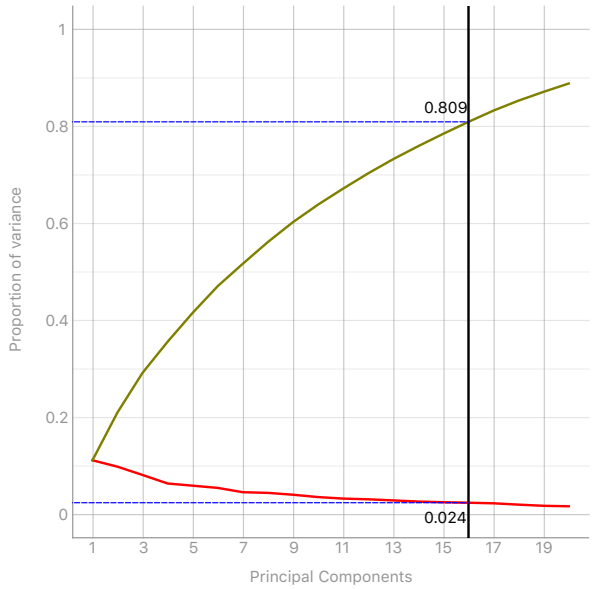
$$u_1^\intercal \mathrm{S} u_1 = u_1^\intercal \lambda_1 u_1 = \lambda_1 u_1^\intercal u_1 = \lambda_1 \tag{3.8}$$

Na levi je izraz za našo varianco $\mathrm{Var}(u_1^\intercal X^\intercal)$, ki jo skušamo maksimizirati. Vemo, da mora biti

u_1 lastni vektor kovariančne matrike. Skladno z zgornjo enačbo pa sedaj tudi vemo, da je to lastni vektor z najvišjo pripadajočo lastno vrednostjo.

V smeri vektorja u_1 se lega naših primerov X v n-dimenzionalnem prostoru torej najbolj spreminja, njihova varianca je v tej smeri največja. Z $u_1^\intercal X^\intercal$ dobimo pravokotno projekcijo vsakega od primerov na to os. S to projekcijo, oziroma legi v smeri u_1, smo pojasnili največjo varianco primerov. Kaj še ostane? Vse, kar je pravokotno na smer u_1. Največjo varianco na pravokotni hiperravnini pa je ravno v smeri drugega lastnega vektorja, saj je ta pravokoten na u_1, to je vektorja, ki ima drugo največjo lastno vrednost. Pojasnjen delež te variance je λ_2, skupaj pa s pravokotno projekcijo na prvi in drugi lastni vektor pojasnimo $\lambda_1 + \lambda_2$ variance množice primerov.

Lastni vektorji kovariančne matrike nam določajo nov koordinatni sistem, kjer so koordinate, po vrsti, tiste, po katerih se lege točk v večdimenzionalnem prostoru najbolj spreminjajo, oziroma je njihova varianca največja. Nove koordinate, urejene skladno z lastnimi vrednostmi vektorjev, imenujemo komponente primerov v novem, transformiranem sistemu. Ker nas bodo zanimale samo te z visokimi deleži razložene variance jih bomo imenovali *glavne komponente*.



Slika 3.3: Graf odvisnosti razložene variance od števila glavnih komponent (zgornja črta) za podatke o torbicah iz slike 3.1.

V splošnem nas zanima, koliko najvišje rangiranih dimenzij v novem koordinatnem sistemu zares potrebujemo za opis podatkov. Pri tem moramo seveda določiti, kakšen je naš ciljni delež pojasnjene variance. Tipično smo zadovoljni, če izbrano število komponent pojasni vsaj 80% skupne variance. Ker skupno varianco poznamo (enaka je vsoti kvadratov razdalj do centra), je potrebno torej le določiti, koliko začetnih urejenih lastnih vrednosti moramo

sešteti, da njihova vsota predstavlja želeni delež razložene variance. Primer grafa, ki kaže odvisno deleža razložene variance glede na število glavnih komponent kaže slika 3.3.

Pristop glavnih komponent se mnogokrat uporablja tudi samo za namene vizualizacije, kjer primere, opisane z mnogimi atributi, želimo predstaviti v dvodimenzionalni ravnini. Pri tem se moramo seveda zavedati, da prvi dve komponenti razložijo morda le majhen del celotne variance.

Za konec, naj ponovimo. Za predstavitev podatkov z glavnimi komponentami je potrebno podatke najprej osrediščiti in nato poiskati kovariančno matriko. Lastni vektorji kovariančne matrike so bazni vektorji novega, transformiranega sistema, njihove lastne vrednosti pa povedo, kakšen delež variance nam razloži posamezna komponenta.

3.2 Računski postopki določanja glavnih komponent

Še praktičen razmislek: v primerih velikega števila atributov bo računanje vseh lastnih vektorjev in vrednosti zamudno. Še posebej, če nas zanima samo projekcija podatkov v dvodimenzionalni prostor. Prvi lastni vektor kovariančne matrike M lahko rajši (hitreje) določimo numerično, s potenčno metodo. Vzamemo nek naključni vektor (npr. x), ga transformiramo z matriko M (torej, izračunamo $x \leftarrow Mx$), in to ponavljamo do konvergence. Ob vsakem koraku tako dobljeni x normaliziramo, torej, $x \leftarrow \frac{Mx}{\|Mx\|}$. Na ta način dobimo lastni vektor. Kako izračunamo pripadajočo lastno vrednost?

$$Mu = \lambda u \tag{3.9}$$

$$u^\intercal Mu = u^\intercal \lambda u = \lambda u^\intercal u = \lambda \tag{3.10}$$

Kako določimo naslednjo komponento? Ena od možnosti je, da uporabimo ravnokar izračunani lastni vektor, nanj projiciramo podatke, te vrednosti odštejemo od podatkov (torej dobimo projekcijo na lastnemu vektorju pravokotno hiperravnino) in za te podatke ponovno izračunamo kovariančno matriko ter s potenčno metodo njej lastni vektor z največjo lastno vrednostjo. V primeru, da potrebujemo več komponent, postopek ustrezno ponovimo.

Za primer, ko nas zanima samo projekcija podatkov v ravnino lahko potenčno metodo namesto na vektorju izvedemo na sistemu dveh vektorjev U. Vektorja matrike U morata biti pravokotna. To dosežemo tako, da po vsakem koraku potenčne metode uporabimo Gram-Schmidtovo ortogonalizacijo.