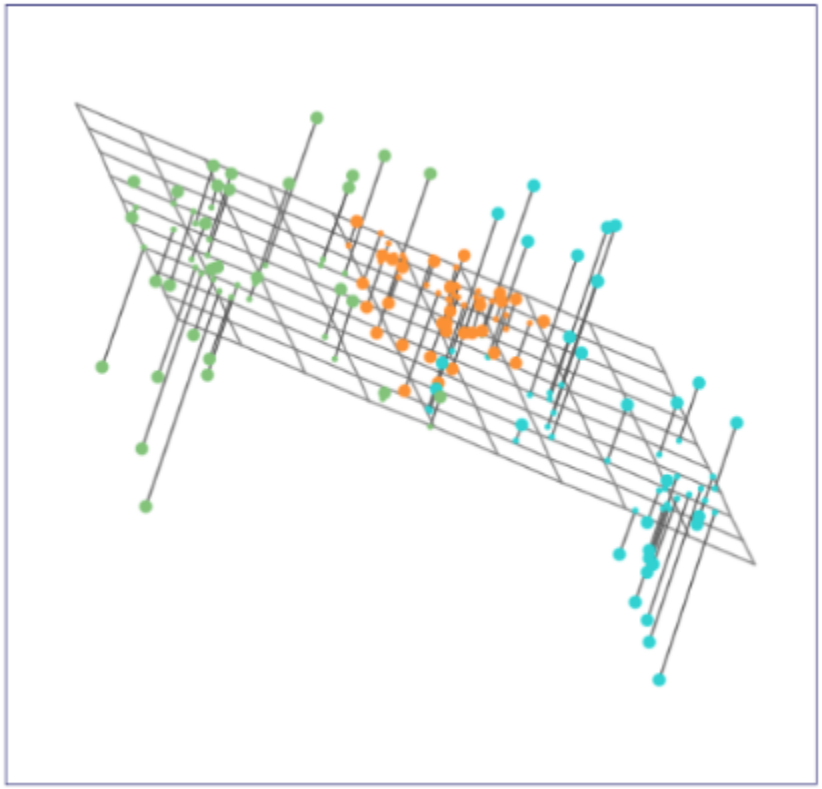
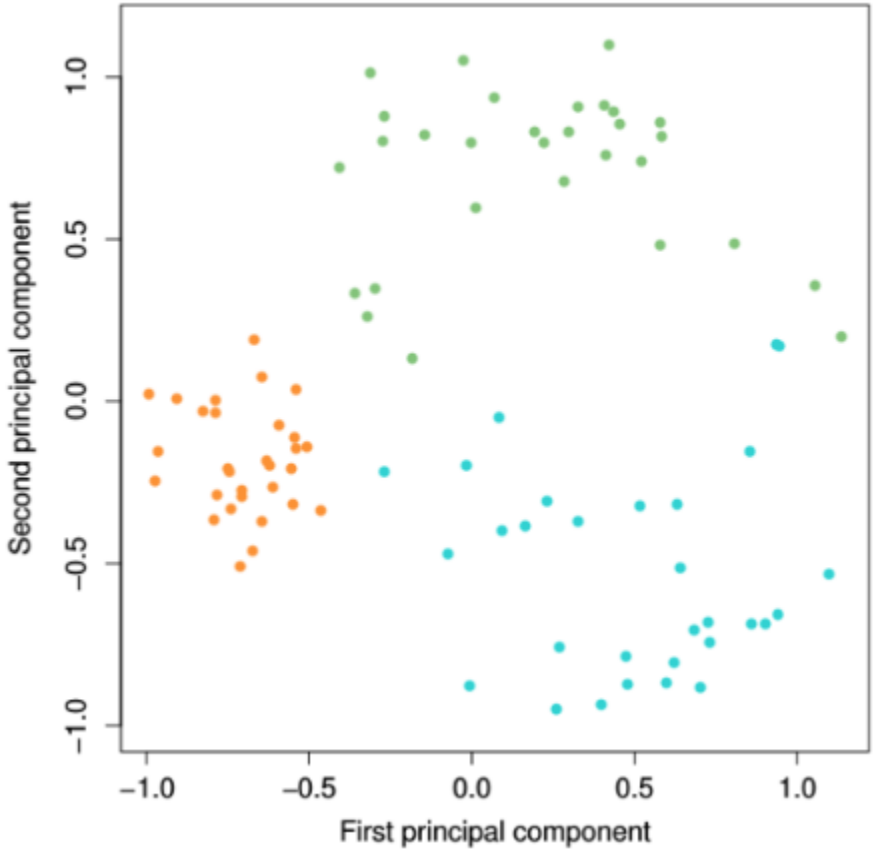




Redukcija dimenzionalnosti

PCA (*Principal Component Analysis*)

Kako da smanjimo broj dimenzija?

- Ideja: kreirati novi podskup obeležja koji dobro sumarizuje polazna obeležja

- Dobar podskup obeležja je onaj koji je *relevantan* za ciljnu funkciju f

- Na primer, onaj koji ima veliki kapacitet da napravi razliku između primera različitih klasa

Zbog čega želimo manje dimenzija?

1. Kompresija
   - manje zauzeće memorije i diska
   - (važnije) značajno ubrzanje obučavajućih algoritama
     - Npr., kompleksnost jednostavnih algoritama poput *k*-means eksponencijalno raste sa dimenzionalnošću

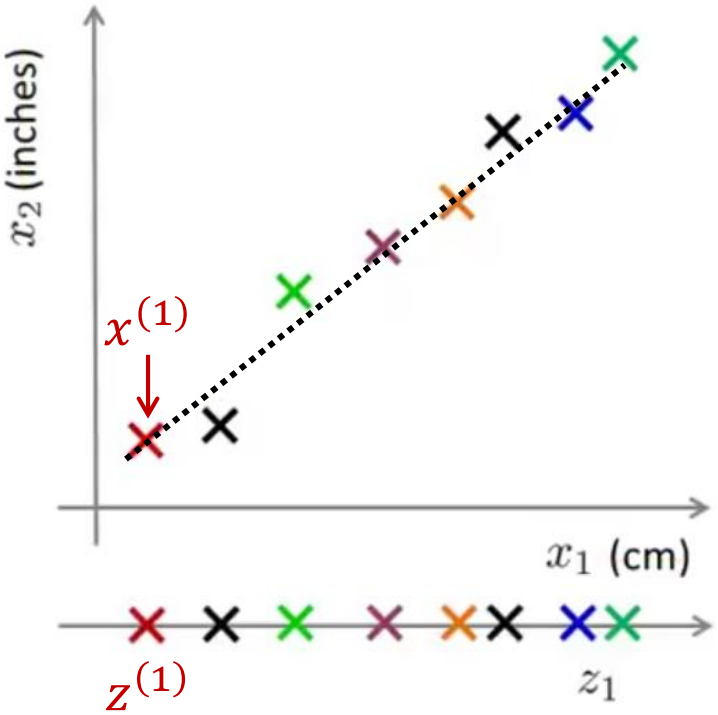
2. Uklanjanje šuma
   - Previše (irelevantnih) obeležja može da degradira performanse

3. Vizuelizacija
   - Bolje razumevanje podataka što može da omogući izgradnju boljih modela

Kompresija

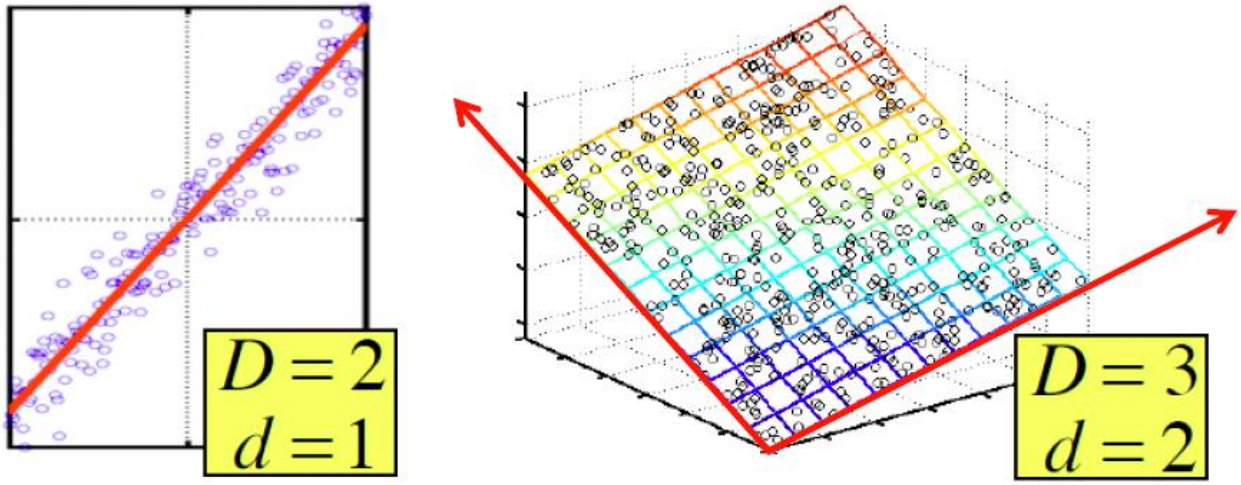
- Recimo da smo sakupili skup podataka sa veoma mnogo obeležja
- Ovde su grafički predstavljena samo dva obeležja:

  x_1 – dužina u cm, x_2 – ista dužinu u inčima



- Umesto da imamo dva odvojena (redudantna) obeležja, bolje bi bilo da redukujemo informaciju u jedno obeležje (jednu dimenziju)
  - $x^{(i)} \in \mathbb{R}^2 \rightarrow z^{(i)} \in \mathbb{R}$
  - Izvršili smo određenu aproksimaciju skupa podataka, ali smo prepolovili broj obeležja

Kompresija



- Pretpostavka: podaci leže tačno na ili blizu *d*-dimenzionog potprostora
- Ose ovog potprostora predstavljaju efektivnu reprezentaciju podataka
- U tipičnom zadatku redukcije dimenzionalnosti možemo imati više hiljada obeležja koja želimo da projektujemo u 100-dimenzioni prostor

Uklanjanje šuma

- Zamislite skup podataka koji se sastoji od dva primera:

$$x^{(1)} = [-1, a_1, a_2, \dots, a_d], y^{(1)} = +1$$
$$x^{(2)} = [+1, b_1, b_2, \dots, b_d], y^{(2)} = -1$$

  gde su $a_i, b_i \in \{-1, +1\}$ slučajne promenljive

- Neka je samo prva komponenta relevantna za ciljnu funkciju f

- Test primer: $x = [-1, -1, \dots, -1]$

- Ako primenjujemo K-NN: korektna klasifikacija zavisi od toga da imamo više -1 među a_i nego među b_i

Uklanjanje šuma

- Još jedan primer bi bilo automatsko prepoznavanje osobe koja se nalazi na slici
  - Interesuju nas sistematične varijacije koje zaista reprezentuju kako osoba izgleda
  - Ali na slikama možemo imati „šum“ poput promena u osvetljenju i drugih uslova pod kojim je snimak napravljen

- Prilikom automatskog klasifikovanja rukom pisanih cifara:
  - Pretvaranje slike u binarne
  - Skaliranje na istu dimenziju, npr. 16×16
  - Umesto 256 parametara možda možemo koristiti svega dva relevantna obeležja – prosečnan intenzitet i simetrija

  …uklanjamo fluktuacije nisu relevantne za prepoznavanje o kojoj je cifri reč

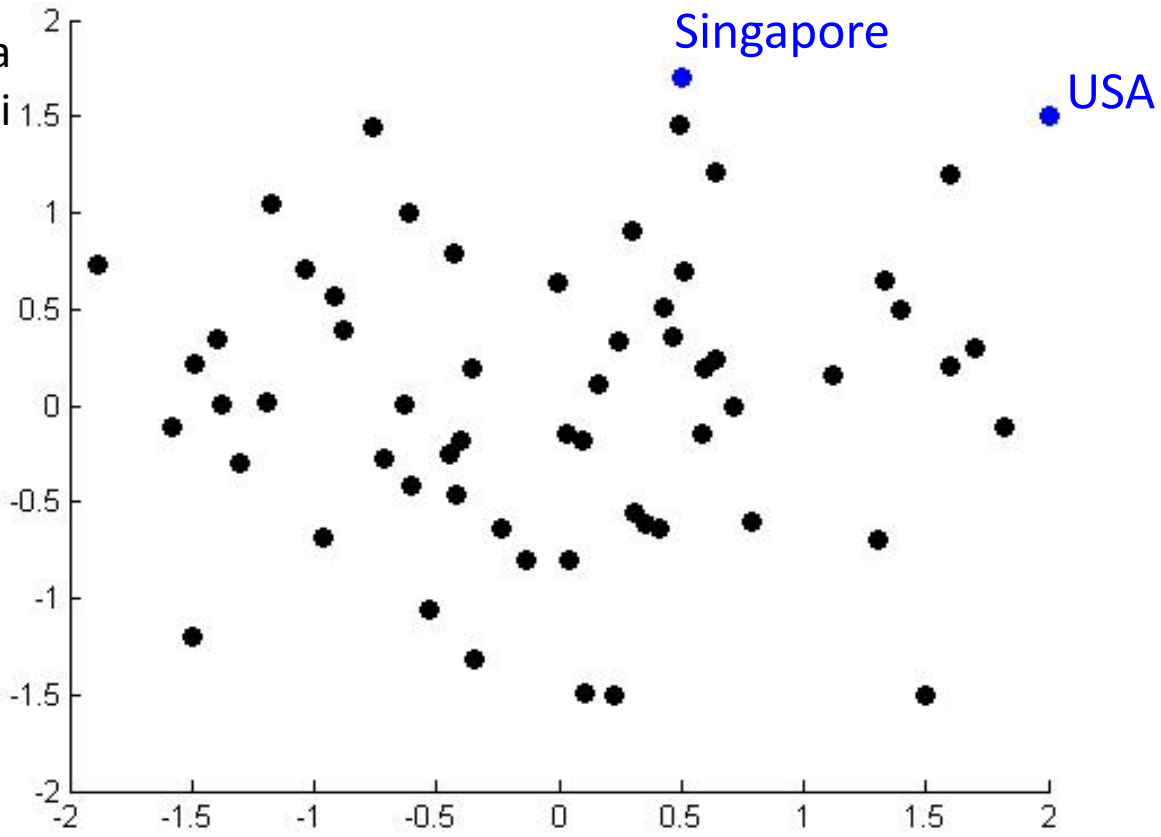
Vizuelizacija

Country	GDP (trillions of US$)	Per capita GDP (thousands of intl. $)	Human Develop-ment Index	Life expectancy	Poverty Index (Gini as percentage)	Mean household income (thousands of US$)	...
Canada	1.577	39.17	0.908	80.7	32.6	67.293	...
China	5.878	7.54	0.687	73	46.9	10.22	...
India	1.632	3.41	0.547	64.7	36.8	0.735	...
Russia	1.48	19.84	0.755	65.5	39.9	0.72	...
Singapore	0.223	56.69	0.866	80	42.5	67.1	...
USA	14.527	46.86	0.91	78.3	40.8	84.3	...
...	...	...	...	...	...	...	

Vizuelizacija

~ GDP per capita/ekonomska aktivnost po osobi

Country		
Canada	1.6	1.2
China	1.7	0.3
India	1.6	0.2
Russia	1.4	0.5
Singapore	0.5	1.7
USA	2	1.5
…	…	…



~ ekonomska aktivnost zemlje (zavisi od veličine zemlje)

Kako da smanjimo broj dimenzija?

- Neka su dati ulazi x. Od njih ćemo konstruisati nove ulaze z primenom neke transformacije:

$$z = \Phi(x)$$

  Ako je dimenzija z manja od dimenzije x, onda smo postigli redukciju dimenzionalnosti

- Idealno, dobili bismo jedno obeležje koje bi bila sama ciljna funkcija $z = f(x)$ - ako bismo imali ovakvo obeležje, naš zadatak je završen
- Ovo sugeriše da je kvalitetna redukcija obeležja jednako teška kao i originalni obučavajući problem pronalaženja ciljne funkcije f

Transformacije obeležja

- Do sada smo mnogo puta videli transformacije obeležja
- U toj postavci, povećanje dimenzionalnosti je bio pokušaj da smanjimo E_{train}, ali, plaćali smo cenu lošije generalizacije (veća razlika između E_{test} i E_{train})
- Ako možemo da smanjimo dimenzionalnost, bez da povredimo E_{train}, pobojšali bismo generalizaciju

Oprez!

- Važno je redukciju dimenzionalnosti sprovesti na principijelan način
- Odbacujemo informacije – možemo da izgubimo one koje su ključne za obučavanje
- Važno je da algoritam sačuva koristan deo informacija, a odbaci šum

Kako da smanjimo broj dimenzija?

- Selekcija obeležja
  - Pronaći minimalan podskup obeležja koji nam može pomoći da razlikujemo klase
- Redukcija dimenzionalnosti
  - Kreirati nova obeležja koja će predstavljati neku kombinaciju starih obeležja