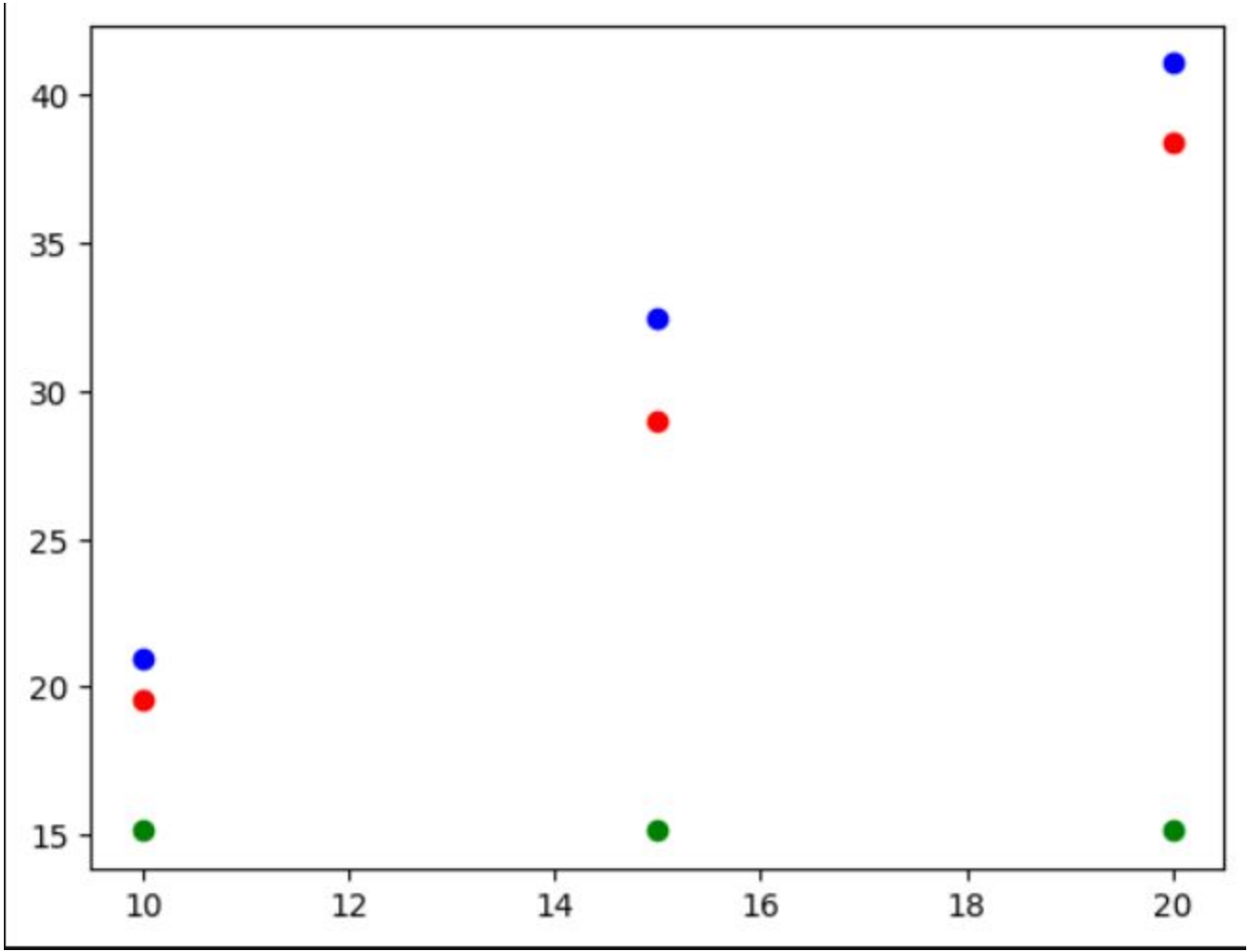
Vektorizacija sadržaja

Šta smo radili?

- ML - skup algoritama, učenje obrazaca
- Supervised, unsupervised i RL
- Regresija = kontinualne vrijednosti
- Klasifikacija = diskretna vrijednost iz predefinisanog skupa
- Trening vs test skup
- Regresija i numpy.polyfit
- Overfitting vs underfitting
- Klasifikacija i kNN algoritam
- Evaluacija i zašto tačnost nije dobra?

Zadatak 1?

- Kako prilagoditi KNN za regresiju? - umjesto da prebrojimo koja se klasa najčešće javlja, nađemo srednju vrijednost rezultata
- Koji su problemi?
  - Svi koji već važe sa KNN - memorija i brzina
  - skup podataka mora biti reprezentativan za dati problem
- Linearna regresija sa KNN - kako se predviđaju vrijednosti koje su van trening skupa (recimo da trening skup ima tačke između 0 i 10 a da nas zanima izlaz za tačku 15) ?


40
35
30
25
20
15
10
12
14
16
18
20

Koje su bolje metrike od tačnosti?

- Preciznost i tačnost u različite - Accuracy (tačnost)
- Matrica konfuzije (confusion matrix) - razlikujemo 4 tipa rezultata predikcije binarnog klasifikatora:
  - tačno pozitivni (True Positive - TP)
  - tačno negativni (True Negative - TN)
  - lažno pozitivni (False Positive - FP)
  - lažno negativni (False Negative - FN)
- Idealan klasifikator = samo TP i TN

Nove metrike iz matrice konfuzije

- Tačnost (Accuracy) = (TP + TN) / (TP + TN + FP + FN)
- Preciznost (Precision) = TP / (TP + FP)
- Preciznost mjeri koliko smo tačno predvidjeli pozitivne događaje
- Odziv (Recall) = TP / (TP + FN)
- Odziv mjeri koji procenat pozitivnih dogđaja predviđamo
- Težnja da se sistem opiše jednim brojem (jednom metrikom) - idealno kombinacija rezultata metrike preciznosti i odziva
- F1 i AUC metrike

F1 vs AUC

- F1 = 2 * (Precision * Recall) / (Precision + Recall)
- F1 predstavlja harmonijsku sredinu preciznosti i odziva - niska vrijednost bilo koje od dve metrike značajno smanjuje vrijednost F1 metrike
- AUC ili češće AUROC (Area Under the ROC (Receiver Operating Characteristic)) - vjerovatnoća da će nasumično odabrana pozitivna klasa imati veću vrijednost od nasumično odabrane negativne klase

Granica odlučivanja

- Linija (2D) koja razdvaja regione različitih predviđenih klasa
- Sve tačke na jednoj strani granice su klasifikovane kao jedna klasa a na drugoj kao druga klasa
- Vizualizacija rada klasifikatora
- Oblik i složenost granice nam može ukazati na performanse klasifikatora
  - veoma uvijena granica može ukazati na overfitting
  - prefiše glatka granica - underfitting
- Kako k utiče na granicu odlučivanja u KNN algoritmu?

Vektorizacija sadržaja

- Kako bismo koristili algoritme mašinskog učenja svaki ulaz mora biti pretvoren u vektor - brojeve
- Većina podataka nije u brojevnom formatu - slike, tekst, zvuk, datumi rođenja...
- Bitno definisati jasne i konzistentne načine konverzije ulaznog sadržaja u vektore
- Način se bira u zavisnosti od problema, ali postoje poznate tehnike koje se često koriste za odgovarajuće forme ulaza

Kategorički podaci

- Neizbježan većine dataset-ova (pol, zanimanje, nivo stručne spreme...)
- Dve najčešće tehnike
- Prva tehnika - kategorija kao redni broj
  - Svakoj klasi dodijelimo odgovarajući broj počevši od nule
  - Jednostavna implementacija
  - Niska dimenzionalnost - kategorija postaje integer, tako da se ne povećavaju memorijski zahtjevi ili smanjuje brzina
- Druga tehnika - One-hot encoding

One-hot encoding

- Koraci:
  - Konverzija kategorije u redni broj
  - Kreiranje vektora popunjenog nulama čija je dimenzija jednaka broju kategorija
  - Postavljanje jedinice na mjesto koje odgovara rednom broju trenutne kategorije
- Prednosti - nema implicitne ordinalne zavisnosti - u slučaju KNN i linearne regresije može pogoršati performanse
- Nedostaci - memorija, brzina

Pretraga teksta

- Rad sa tekstom postao veoma aktuelan - ChatGPT i drugi veliki jezički modeli
- Kompleksan problem, mi ćemo rješavati malo jednostavniji
- Kako naći odgovarajuće tekstualne dokumente na osnovu upita ili nekog drugog tekstualnog dokumenta
- Najjednostavnije - da li baš zadate riječi postoje u dokumentu - veoma loše performanse
- Prvi korak i dalje isti - pretvoriti tekst u vektor
- Fokus - metode zasnovane na vektorskom prostoru

Vektorizacija teksta

- Bag of words - tekst kao skup riječi
- Problem = riječi u različitim formama (padeži, vremena, lica...)
- Bitne tehnike:
  - uklanjanje “stop words” - riječi koje se često javljaju i nemaju neko značenje (a, the, is, and...)
  - steming - svođenje riječi na korijen (uklanjamo sufikse/prefikse)
  - lematizacija - naprednija tehnika, uzima u obzir gramatičku strukturu, upotrebu i druge aspekte

Normalizovali smo tekst, šta dalje?

- Nakon normalizacije teksta potrebno ga je pretvoriti u vektor
- Bitna ideja - tekst je skup riječi
- Rječnik - skup svih riječi u našim dokumentima
- Dva osnovna pristupa
  - Count Vectorizer - koliko se svaki token iz našeg rječnika pojavljuje u našem tekstu
  - TF-IDF Vectorizer - malo pametniji pristup
- Za oba pristupa postoje gotove klase

TF-IDF

- Problem sa prebrojavanjem
  - ne uzima u obzir značaj neke riječi
  - veliki dokumenti - više riječi - veći brojevi
- TF-IDF - dva faktora za svaku riječ
  - TF (Term Frequency) - frekvencija pojavljivanja tokena t u dokumentu d
  - IDF - koliko je riječ značajna za cijeli korpus teksta, koliko se često token t pojavljuje u cijelom korpusu D

TD-IDF

- TF(t, d) = (broj pojavljivanja tokena t u dokumentu d)/(broj tokena u dokumentu d)
- IDF(t,D) = log ((broj dokumenata u korpusu D)/(broj dokumenata koji imaju token t))
- TF-IDF(t, d, D) = TF(t, d) * IDF(t, D)
- Ideja = tokeni koji se često pojavljuju u nekom dokumentu, a veoma rijetko u drugima imaju veliki značaj (razlikuju dati dokument od ostalih)

Pretraga teksta

- Trebaju nam dokumenti koji su najsličniji nekom upitu
- Upit može biti slobodan tekst ili neki drugi dokument
- Imamo već pripremljen tekst iz našeg korpusa
- Bitno - obraditi ulazni tekst na isti način kao tekst iz korpusa
- KNN algoritam za pretragu
- Ne tražimo više klasu koja se najčešće pojavljuje, samo vraćamo K najbližij susjeda

Zaključak

- Matrica konfuzije - preciznost i odziv bolje metrike
- KNN za regresiju - pronaći srednju vrijednost, koji su problemi
- Žašto nam je bitna vektorizacija?
- Kategorički podaci - one hot encoding
- Rad sa tekstom
- Preprocessing teksta - stemming, lematizacija, stop words
- Vektorizacija teksta - Count, TF-IDF
- Pretraga teksta - KNN koraci

Hvala na pažnji!