Avaluació experimental de l'algorisme k-means i variants

Professorat d'Algorísmia (GRAU-A)
Departament de Ciències de la Computació
Universitat Politècnica de Catalunya

Q2 2023–2024

1 Objectiu general i normes

Aquest projecte té com a objectiu la implementació de l'algorisme de clustering anomenat k-means (almenys dues de les seves variants) i el seu estudi experimental, tant de la seva eficiència en temps com de la qualitat de les solucions obtingudes.

El projecte es farà en grups de **4 persones**. Per formalitzar els grups us heu d'apuntar al fitxer compartit `Equips Projecte GRAU-A Q2 2023-24` (seguiu l'enllaç). En aquest fitxer trobareu una columna per introduir un identificador d'equip (1, 2, 3, ...) i a continuació, columnes per a introduir els cognoms, nom i subgrup de cadascun dels integrats de l'equip, seguint el model d'exemple. Això s'ha de fer **abans del 26 de febrer de 2024**. Els estudiants que no hagin format grup fins aquesta data, seran organitzats en equips pel professorat de l'assignatura.

El lliurament dels materials demanats en aquest projecte es farà en línia via el **Racó FIB**. La data límit d'entrega són les 23:59 hores del dia **22 de març de 2024**.

En qualsevol moment durant el procés de correcció podríeu ser contactats per part d'algun professor de l'assignatura per tal de resoldre dubtes o fer aclariments sobre el vostre treball.

Totes les comunicacions públiques referents al projecte es duran a terme mitjançant el *Racó FIB* o el *canal de Slack* `#projecte`.

2 Especificació del projecte

Els continguts d'aquest projecte s'organitzen en tres parts:

A la **primera part** implementareu variants de l'algorisme k-means, o d'algun altre algorisme semblant com k-mediods o k-medians, l'objectiu dels quals és subdividir un conjunt d'n elements en k classes o clústers, on cada clúster agrupa objectes semblants entre si, és a dir, a poca distància els uns dels altres. S'han d'implementar dos algorismes com a mínim (entenent cada variant de l'algorisme k-means com un algorisme diferent).

La **segona part** està centrada a mesurar la qualitat de les particions obtingudes pels nostres algorismes implementats a la primera part. Implementareu, almenys, **(a)** una mesura de qualitat *interna*, és a dir, una mesura que dona una puntuació o valor a una partició donada del conjunt d'entrada; i **(b)** una mesura de qualitat *externa*, és a dir, una mesura de la similitud (o no) entre dues particions d'un mateix conjunt, la qual cosa ens permet comparar solucions alternatives obtingudes mitjançant algorismes diferents, o comparar una partició amb la partició de referència, la *groundtruth*. També dintre d'aquesta segona part implementareu almenys un mètode, d'entre els molts que s'han proposat, que ens permeti determinar el valor òptim de k.

La **tercera part** es dedicarà a l'anàlisi experimental pròpiament dita, amb experiments amb els quals mesurareu l'eficiència en temps d'execució dels diferents algorismes de *clustering* implementats en la primera part, la qualitat de les solucions obtingudes, usant les mesures internes i les mesures externes per a fer comparatives entre els diferents algorismes, i aplicareu el o els mètodes de determinació de la k òptima.

Part 1: Algorisme de k-means

L'algorisme de k-means estàndard és un algorisme iteratiu que va refinant una partició inicial del conjunt d'n elements d'entrada en k clústers. Cada clúster té un *centre* (o *centroide*) que és la mitjana aritmètica dels elements que formen part del clúster, considerats aquests com a punt d-dimensionals. Durant una iteració, cadascun dels n elements s'assigna al centre més proper (a distància euclidiana mínima) i en acabar s'actualitzen els centres dels k clústers. L'algorisme acaba quan s'arriba a una determinada condició (p. ex., nombre d'iteracions prefixat o quan la partició no canvia o quasi d'una iteració a la següent).

En funció de com definim els centres dels clústers, de com es fa la partició inicial, o de quan detenim l'execució de l'algorisme tenim moltes variants que s'han definit i estudiat àmpliament en la literatura. Per exemple, per a la inicialització podem escollir k elements del conjunt a l'atzar i formar els k clústers inicials assignant cada element del conjunt a l'element més proper d'entre els k escollits, és a dir, apliquem una iteració de l'algorisme com si els k elements escollits fossin els centres de k clústers inicials. Una altra famosa variant és k-means++ per a obtenir una partició inicial bona.

Depenent de com calculem el centre d'un clúster tenim variacions de l'esquema bàsic de k-means com ara k-medians o k-mediods.

Implementeu almenys dues variants de l'algorisme de k-means; us aconsellem implementar com a mínim la versió més bàsica, coneguda també com l'algorisme de Lloyd.

Part 2: Mètodes d'avaluació de clusterings i d'optimització de k

En aquesta part haureu d'implementar tres tipus d'algorismes per a l'estudi comparatiu de la qualitat dels algorismes de clustering.

Específicament haureu d'implementar almenys una mesura interna de qualitat, que atorga una puntuació a una partició donada. Per exemple, la qualitat de la partició la podem mesurar calculant la ràtio entre les distàncies mínima i màxima entre els centres dels k clústers o calculant la distància *intra-clúster* (= distància mitjana entre elements d'un clúster) i després fer mitjana de les distàncies intra-clúster. A la literatura s'han definit molts altres coeficients que mesuren el grau de "consistència" d'una partició, com ara l'índex de Calinski–Harabasz, l'índex Davies-Boudin, o la silueta mitjana (*average silhoutte*).

També heu d'implementar almenys una mesura externa de qualitat. En una mesura externa de qualitat comparem dues particions d'un mateix conjunt. És, fonamentalment, una mesura de similitud entre dues particions del mateix conjunt. Una de les més conegudes és el *Rand Index*. Per a cada parell (x, y) d'elements, i les dues particions A i B, es diu que A i B concorden si A i B classifiquen x i y de la mateixa manera, és a dir, si x i y són a un mateix clúster en A llavors també són en un mateix clúster en B, i si x i y són a clústers diferents en A també són a clústers diferents en B. Dividint el nombre de parells concordants entre el total de parells $\binom{n}{2}$ s'obté el *Rand Index*. Noteu que no es posen en correspondència els clústers d'A amb els de B (fins i tot poden tenir un nombre de clústers diferents). Novament, hi ha moltes altres mesures externes de qualitat, que es poden consultar a l'àmplia literatura existent.

Finalment, dintre d'aquesta part hauríeu d'implementar almenys un mètode per a determinar el nombre de clústers. De vegades aquest nombre és conegut degut al nostre coneixement del domini i la provinença del conjunt de dades, però altres vegades no és així. Si el valor de k és molt baix ($k = 1$ en cas extrem) la partició obtinguda no serà útil i no tindrà poder explicatiu ni predictiu quan utilitzem el model generat per l'algorisme de clustering. A l'extrem oposat un valor k molt elevat dona clústers amb molt bona distància intra-clúster perquè al voltant de cada centre hi haurà pocs elements en cada clúster, però tampoc això ens serveix. Es tracta de trobar un valor de k que ens doni clústers "compactes" i que no siguin més dels necessaris. S'han proposat diversos mètodes per a trobar el millor valor de k, com per exemple el mètode del colze (*elbow method*). Aquest mètode sovint es presenta de manera gràfica (i vosaltres haureu de fer aquestes gràfiques en el vostre estudi), però automatitzeu el seu càlcul. Per a aplicar el mètode del colze, l'algorisme de clustering s'executa amb valors de k successius i es determina el valor òptim de k quan la variació en la mesura interna de qualitat escollida, passant d'un valor de k al següent, és petita.

Part 3: Experimentació

L'objectiu d'aquesta part del projecte és portar a terme experiments en els quals executareu els algorismes de la part 1 sobre un seguit de conjunts de dades (*datasets*), obtenint en cada experiment un clustering del *dataset* corresponent, del qual s'obtindran les mesures internes de qualitat i es compararà amb els clustering obtinguts amb altres algorismes alternatius i també amb la *groundtruth* si les dades del *dataset* venen etiquetades amb els clústers "reals". També s'haurà mesurat el temps d'execució de cada algorisme amb cada *dataset*.

Com que algunes de les variants utilitzen al·leatorietat pel seu funcionament alguns experiments s'hauran d'executar múltiples cops per obtenir-ne valors mitjans. Això val per a tots els algorismes, inclús quan són deterministes, de cara a les mesures de temps d'execució, repetint l'experiment obtindreu estimacions més robustes del temps d'execució. Com cada execució independent de l'algorisme amb el mateix *dataset* ens pot conduir a una partició diferent de les dades, es farà la mitjana de les mesures internes, de les mesures externes, etc.

En un altre grup d'experiment aplicareu el mètode de colze o el mètode implementat per a determinar experimentalment el valor òptim de k.

Mitjançant Slack us donarem accés als conjunts de dades que heu d'utilitzar com a entrada dels vostres algorismes en els experiments. Cada conjunt de dades està en un fitxer de text i comença indicant el nombre n d'elements del *dataset* i la dimensionalitat d de cada element. També es dona el nombre de clústers. A continuació venen les dades, un element per "línia" consistent en d valors reals (o enters en alguns casos); en alguns *datasets*, i així ho indicarem, cada element està etiqueta amb un identificador de clúster. Els algorismes desenvolupats no han de tenir en compte aquesta informació de cap manera; només es tindrà en compte per a calcular una mesura externa de qualitat del clustering, comparant el que han fet els algorismes respecte a la partició "correcta".

En els experiments de determinació de la k òptima el mateix algorisme s'executarà sobre el *dataset* amb successius valors de k i es tracta de veure que, efectivament, el mètode de colze o alternatiu que hàgim escollit ens determina correctament el valor de k (que en realitat sí és conegut).

3 Avaluació del projecte

Aquest document és intencionadament vague i s'espera que investigueu pel vostre compte totes les tècniques algorísmiques i models que es mencionen aquí. Hi ha molta bibliografia accessible

al respecte i no us costarà gens trobar-ne informació.

Un cop portats a terme els experiments demanats i reunides totes les dades experimentals, heu de preparar un informe en què es descrigui breument la vostra implementació dels algorismes de k-means (descriviu la part comuna, d'una banda, i després les variants, no cal repetir explicacions redundants), dels algorismes necessaris per a avaluar la qualitat de les solucions obtingudes, del o dels algorismes per a optimitzar el paràmetre k (nombre de clústers), i del programa o part del programa encarregat de dur a terme els experiments.

Recordeu que us demanem que implementeu almenys dos algorismes de clustering (part 1), almenys una mesura interna de qualitat, almenys una mesura externa de qualitat i almenys un algorisme de determinació de la k òptim, i portar a terme els experiments descrits a la part 3.

Us valorarem positivament que implementeu més algorismes de clustering, o més mesures internes, etc. i feu els corresponents estudis i comparatives (p. ex., no val només implementar una nova mesura interna de qualitat i no fer els experiments i comparar-la amb la o les altres mesures internes implementades o no descriure el que s'ha fet en l'apartat corresponent, no donar referències bibliogràfiques en el seu cas, ...). També es valorarà positivament que dissenyeu nous experiments que intentin explicar el comportament observat en funció de característiques dels datasets (per exemple, si els clústers estan ben separats, la dimensionalitat, etc.) o que desenvolupeu (i documenteu) un generador de dades sintètiques, per exemple.

L'ús de figures i diagrames per a il·lustrar els conceptes i el funcionament dels algorismes és benvingut. Ans al contrari, incloure codi detallat en el cos principal de l'informe s'ha d'evitar; podeu donar codi detallat en apèndixs del vostre informe, si cal, específicament si voleu fer referència a parts específiques del codi desenvolupat.

L'informe també ha de descriure breument la configuració experimental. Proporcioneu taules i gràfics que resumeixin els resultats dels experiments. En particular, heu de donar gràfics que mostrin com el mètode del colze (*elbow method*) o le mètode alternatiu emprat per a determinar el valor òptim de k de cada *dataset*. Eviteu de les gràfiques en 3D.

L'informe ha d'incloure les vostres conclusions derivades de l'estudi experimental. També és imprescindible que inclogueu apartats descrivint la metodologia de treball, organització de l'equip, valoració del procés d'autoaprenentatge i finalment la bibliografia consultada. És molt fàcil trobar tota mena de material, incloent-hi codi font, relacionat amb els algorismes de clustering i la seva validació/avaluació, si en feu ús és important entendre bé aquests materials i és fonamental referenciar acuradament les fonts.

Us animem a utilitzar LaTeX per preparar el vostre informe. Per fer gràfiques podeu utilitzar qualsevol dels múltiples paquets que té LaTeX (en en particular, els paquets `tikz` i `pgf`) o utilitzar programari independent com ara gnuplot i després incloure en l'informe les imatges/gràfics generades externament.

El nivell de sofisticació i esforç dedicat a la pràctica és opcional i es tindrà en compte a l'hora d'avaluar-la. Tingueu en compte que la documentació entregada ens ha de permetre valorar el nivell d'assoliment de la competència transversal *Capacitat d'autoaprenentatge* que també avaluem amb el projecte. En el context del projecte hi ha molts aspectes rellevants relacionats amb aquesta competència: des de l'estudi de noves tècniques i models algorísmics, fins al disseny i anàlisi d'experiments, i la documentació d'aquests tipus de treballs de recerca.

La qualificació final del projecte reflectirà la qualitat del vostre aprenentatge, de l'experimentació feta i de la documentació lliurada. La qualitat del codi entregat (programes) es pressuposa i representarà una part molt petita de la qualificació final.

4 Detalls de l'entrega

Caldrà lliurar un fitxer comprimit (`.zip`), mitjançant el *Racó*, que contingui el codi font de tots els programes desenvolupats, i tots els fitxers addicionals que puguin ser necessaris per a la compilació i execució de cadascun d'ells, així com les instruccions per fer-los anar. A part dels programes, s'ha d'incloure l'informe del projecte en PDF. Anomeneu aquest fitxer `InformeEquipID.pdf` (on `ID` és l'identificador del vostre equip de treball). No cal incloure els fitxers d'entrada utilitzats en els experiments, però convé incloure els fitxers amb totes les dades experimentals obtingudes.