

En Yakın k-Komsu (k-Nearest Neighbor)

Yapay Ogrenim alanında örnek bazlı öğrenen algoritmalarından bilinen kNN, eğitim verinin kendisini sınıflama (classification) amaçlı olarak kullanır, yeni bir model ortaya çıkartmaz. Algoritma şöyle işler: etiketleri bilinen eğitim verisi alınır ve bir kenarda tutulur. Yeni bir veri noktası sorgulunca bu veriye geri donulur ve o noktaya “en yakın” k tane nokta bulunur. Daha sonra bu noktaların etiketlerine bakılır ve çoğunluğun etiketi ne ise, o etiket yeni noktanın etiketi olarak kabul edilir. Mesela elde 1 kategorisi altında $[2 \ 2]$, 2 kategorisi altında $[5 \ 5]$ var ise, yeni nokta $[3, \ 3]$ için yakınlık açısından $[2 \ 2]$ bulunmalı ve etiket olarak 1 sonucu döndürülmelidir.

Ustte tarif edilen basit bir ihtiyaç, yöntem gibi görülebilir. Fakat yapay öğrenim ve yapay zeka çok boyutlarda oruntu tanıma (pattern recognition) ile uğraşır, ve milyonlarca satırlık veri, onlarca boyut (ustteki örnekte 2, fakat çoğunlukla çok daha fazla boyut vardır) işler hakikaten zorlaşabilir. Mesela görüntü tanımada veri $M \times N$ boyutundaki dijital imajlar (düzleştirilince $M \cdot N$ boyutunda), ve onların içindeki resimlerin kime ait olduğu etiket bilgisi olabilir. kNN bu tür multimedia, çok boyutlu veri ortamında başarılı şekilde çalışabilmektedir. Ayrıca en yakın k komşunun içeriği tarifsel bilgi çıkarımı (knowledge extraction) amacıyla da kullanılabilir [2].

“En yakın” sözü bir koordinat sistemi anlamına geliyor, ve kNN, aynen k-Means ve diğer pek çok koordinatsal öğrenme yöntemi gibi eldeki çok boyutlu veri noktalarının elemanlarını bir koordinat sistemindeymiş gibi görür. Kiyasla mesela APriori gibi bir algoritma metin bazlı veriyle olduğu gibi çalışabilirdi.

Peki arama bağlamında, bir veri obagi icinden en yakın noktaları bulmanın en basit yolu nedir? Listeyi bastan sonra taramak (kaba kuvvet yöntemi -brute force-) listedeki her nokta ile yeni nokta arasındaki mesafeyi teker teker hesaplayıp en yakın k taneyi icinden seçerdi, bu bir yöntemdir.. Bu basit algoritmanın yuku $O(N)$ 'dir. Eğer tek bir nokta ariyor olsaydik, kabul edilebilir olabilirdi. Fakat genellikle bir sınıflayıcı (classifier) algoritmasının sürekli işlemesi, mesela bir online site için günde milyonlarca kez bazı kararları alması gerekebilir. Bu durumda ve N 'in çok büyük olduğu şartlarda, ustteki hız bile yeterli olmayacaktır.

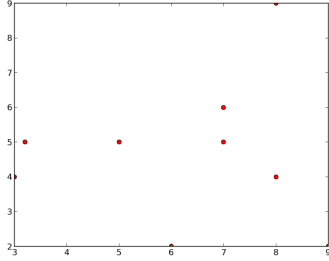
Arama işlemini daha hızlı yapmanın yolları var. Akıllı arama algoritmaları kullanarak eğitim verilerini bir ağac yapısı üzerinden tarayıp erişim hızını $O(\log N)$ 'e indirmek mümkündür.

Küre Ağaçları (Ball Tree, BT)

Bir noktanın diğer noktalara yakın olup olmadığının hesabında yapılması gereken en pahalı işlem nedir? Mesafe hesabıdır. BT algoritmasının puf noktası bu hesabi yapmadan, noktalara değil, noktaları kapsayan “kurelere” bakarak hız kazandırmasıdır. Noktaların her biri yerine o noktaları temsil eden kurenin mihenk noktasına (pivot -bu nokta kure içindeki noktaların ortalamasal olarak merkezi de olabilir, herhangi bir başka nokta da-) bakılır, ve oraya olan mesafeye göre bir kure altındaki noktalara olabilecek en az ve en fazla uzaklık hemen anlaşılmış olur.

Not: Kure kavrami uc boyutta anlamlı tabii ki, iki boyutta bir cemberden bahsetmek lazım, daha yüksek boyutlarda ise merkezi ve çapı olan bir “hiper yüzeyden” bahsetmek lazım. Tarifi kolaylaştırdığı için cember ve kure tanımlarını kullanıyoruz.

Mesela elimizde alttaki gibi noktalar var ve kureyi oluşturduk.

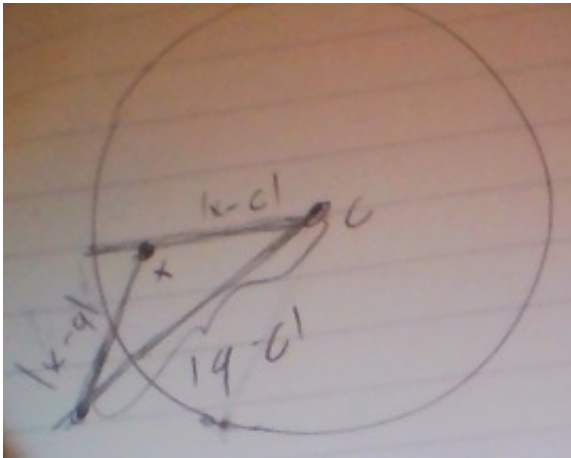


Bu kureyi kullanarak kure dışındaki herhangi bir nokta q 'nın kuredeki “diğer tüm noktalar x 'e” olabileceği en az mesafenin ne olacağını uçgenel eşitsizlik ile anlayabiliriz.

Uçgenel eşitsizlik

$$|x - y| \leq |x - z| + |z - y|$$

$\|$ operatörü norm operatörü anlamına gelir ve uzaklık hesabının genelleştirilmiş halidir. Konu hakkında daha fazla detay için *Fonksiyonel Analiz* ders notlarımıza bakabilirsiniz. Kısaca söylenmek istenen iki nokta arasında direk gitmek yerine yolu uzatırsak, mesafe artacaktır. Tabii uzaklık, yol, nokta gibi kavramlar tamamen soyut matematiksel ortamda da işleyecek şekilde ayarlanmıştır. Mesela mesafe (norm) kavramını değiştirebiliriz, Oklitsel yerine Manhattan mesafesi kullanırız, fakat bu kavram bir norm olduğu ve belirttiğimiz uzayda geçerli olduğu için uçgenel eşitsizlik üzerine kurulmuş tüm diğer kurallar geçerli olur.



Şimdi diyelim ki dışarıdaki bir q noktasından bir kure içindeki diğer tüm x noktalarına olan mesafe hakkında bir şeyler söylemek istiyoruz. Üstteki şekilde bir

ucgensel esitsizlik cikartabiliriz,

$$|x - c| + |x - q| \geq |q - c|$$

Bunun dogru bir ifade oldugunu biliyoruz. Peki simdi yaricapi bu ise dahil edelim, cunku yaricap hesabi bir kere yapilip kure seviyesinde depolanacak ve bir daha hesaplanmasi gerekmeyecek, yani algoritmayi hizlandiracak bir sey olabilir bu, o zaman eger $|x - c|$ yerine yaricapi kullanirsak, esitsizlik hala gecerli olur, sol taraf zaten buyuktu, simdi daha da buyuk olacak,

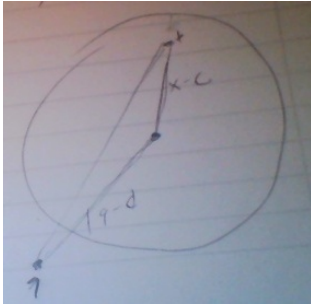
$$\text{radius} + |x - q| \geq |q - c|$$

Bunu nasil boyle kesin bilebiliyoruz? Cunku BT algoritmasi radius'u $|x - c|$ 'ten kesinlikle daha buyuk olacak sekilde secer). Simdi yaricapi saga gecirelim,

$$|x - q| \geq |q - c| - \text{radius}$$

Boylece guzel bir tanim elde ettik. Yeni noktanin kuredaki herhangi bir nokta x 'e olan uzakligi, yeni noktanin mihenke olan uzakliginin yaricapi cikartilmis halinden *muhtak* fazladir. Yani bu cikartma isleminde ele gecen rakam yeni noktanin x 'e uzakligina bir "alt sinir (lower bound)" olarak kabul edilebilir. Diger tum mesafeler bu rakamdan daha buyuk olacaktır. Ne elde ettik? Sadece bir yeni nokta, mihenk ve yaricap kullanarak kuredaki "diger tum noktalar hakkında" bir irdeleme yapmamiz mumkun olacak. Bu noktalara teker teker bakmamiz gerek-meyecek. Bunun nasil ise yaradigini algoritma detaylarinda gorecegiz.

Benzer sekilde



Bu ne diyor?

$$|q - c| + |x - c| \geq |q - x|$$

$|x - c|$ yerine yaricap kullanirsak, sol taraf buyuyecegi icin buyukluk hala buyukluk olarak kalir,

$$|q - c| + \text{radius} \geq |q - x|$$

Ve yine daha genel ve hizli hesaplanan bir kural elde ettik (onceki ifadeye ben-zemesi icin yer duzenlemesi yapalim)

$$|q - x| \leq |q - c| + \text{radius}$$

Bu ifade ne diyor? Yeni noktanin mihenke olan uzakligina yaricap “eklenirse” bu uzakliktan, buyuklukten daha buyuk bir yeni nokta / kure mesafesi olamaz, kuredeki hangi nokta olursa olsun. Bu esitsizlik te bize bir ust sinir (upper bound) vermis oldu.

Algoritma

```

ball_knn( $PS^{in}, node$ )
1   – Eger alttaki sart gecerli ise node icindeki bir noktanin daha once
2   – kesfedilmis k en yakin komsudan daha yakin olmasi imkansizdir
3   if  $D_{minp}^{node} \geq D_{sofar}$ 
4       return  $PS_{in}$  degismemis halde;
5   else if node bir cocuk noktasi ise
6        $PS_{out} = PS_{in}$ ;
7       for  $\forall x \in points(node)$ 
8           if  $(|x - q| < D_{sofar})$ ; – basit lineer arama yap
9            $x$ 'i  $PS_{out}$ 'a ekle;
10          if  $|PS^{out}| == k + 1$ ;
11              en uzak olan komsuyu  $PS^{out}$ 'tan cikart;
12               $D_{sofar}$ 'i guncelle;
13
14   – eger uc nokta degil ise iki cocuk dugumden daha yakin olanini
15   – incele, sonra daha uzakta olanina bak. buyuk bir ihtimalle
16   – arama devam ettirilirse bu arama kendiliginden kesilecektir
17   else
18        $node_1 = node$ 'un  $q$ 'ya en yakin cocugu;
19        $node_2 = node$ 'un  $q$ 'dan en uzak cocugu;
20        $PS^{temp} = ball\_knn(PS^{in}, node_1)$ ;
21        $PS^{out} = ball\_knn(PS^{temp}, node_2)$ ;

```

Kure Agaclari (BT) metodu once kureleri, agaclari olusturmalidir. Bu kureler hiyerarsik sekilde planlanir, tum noktalarin icinde oldugu bir “en ust kure” vardir her kurenin iki tane cocuk kuresi olabilir. Belli bir (disaridan tanimlanan) minimum r_{min} veri noktasina gelinceye kadar sadece noktalar geometrik olarak kapsamakla gorevli kureler olusturulur, kureler noktalar sahiplenmezler. Fakat bu r_{min} sayisina erisince (artik oldukca alttaki) kurelerin uzerine noktalar konacaktır.

Once tek kurenin olusturulusuna bakalim. Bir kure olusumu icin eldeki veri icinden herhangi bir tanesi mihenk olarak kabul edilebilir. Daha sonra bu mihenkten diger tum noktalara olan uzaklik olculur, ve en fazla, en buyuk olan uzaklik yaricap olarak kabul edilir (her sey kapsayabilmesi icin).

Not: Bu arada "tüm diğer noktalara bakılması" dedik, bundan kaçınmaya çalışmıyor muyduk? Fakat dikkat, "küre oluşturulması" evresindeyiz, k tane yakın nokta arama evresinde değiliz. Yapmaya çalıştığımız aramaları hızlandırmak - eğitim / küre oluşturma bir kez yapılacak ve bu eğitilmiş küreler bir kenarda tutulacak ve sürekli aramalar için ardı ardına kullanılacaklar.

Küreyi oluşturma algoritması şöyledir: verilen noktalar içinde herhangi birisi mihenk olarak seçilir. Sonra bu noktadan en uzakta olan nokta f_1 , sonra f_1 'den en uzakta olan nokta f_2 seçilir. Sonra tüm noktalara teker teker bakılır ve f_1 'e yakın olanlar bir gruba, f_2 'ye yakın olanlar bir gruba ayrılır.

```
import itertools

def dist(vect, x):
    return np.fromiter(itertools.imap
                        (np.linalg.norm, vect-x), dtype=np.float)

def norm(x, y): return np.linalg.norm(x-y)

points = np.array([[3., 3.], [2., 2.]])
q = [1., 1.]
print 'diff', points-q
print 'dist', dist(points, q)

diff [[ 2.  2.]
      [ 1.  1.]]
dist [ 2.82842712  1.41421356]

# k-nearest neighbor Ball Tree algorithm in Python
import pprint

__rmin__ = 2

# node: [pivot, radius, points, [child1, child2]]
def new_node(): return [None, None, None, [None, None]]

def zero_if_neg(x):
    if x < 0: return 0
    else: return x

def form_tree(points, node, all_points, plot_tree=False):
    pivot = points[0]
    radius = np.max(dist(points, pivot))
    if plot_tree: plot_circles(pivot, radius, points, all_points)
    node[0] = pivot
    node[1] = radius
    if len(points) <= __rmin__:
        node[2] = points
        return
    idx = np.argmax(dist(points, pivot))
    furthest = points[idx, :]
    idx = np.argmax(dist(points, furthest))
    furthest2 = points[idx, :]
    dist1 = dist(points, furthest)
```

```

dist2=dist(points,furthest2)
diffs = dist1-dist2
p1 = points[diffs <= 0]
p2 = points[diffs > 0]
node[3][0] = new_node() # left child
node[3][1] = new_node() # right child
form_tree(p1,node[3][0],all_points)
form_tree(p2,node[3][1],all_points)

# knn: [min_so_far, [points]]
def search_tree(new_point, knn_matches, node, k):
    pivot = node[0]
    radius = node[1]
    node_points = node[2]
    children = node[3]

    # calculate min distance between new point and pivot
    # it is direct distance minus the radius
    min_dist_new_pt_node = norm(pivot,new_point) - radius

    # if the new pt is inside the circle, its potential minimum
    # distance to a random point inside is zero (hence
    # zero_if_neg). we can only say so much without looking at all
    # points (and if we did, that would defeat the purpose of this
    # algorithm)
    min_dist_new_pt_node = zero_if_neg(min_dist_new_pt_node)

    knn_matches_out = None

    # min is greater than so far
    if min_dist_new_pt_node >= knn_matches[0]:
        # nothing to do
        return knn_matches
    elif node_points != None: # if node is a leaf
        print knn_matches_out
        knn_matches_out = knn_matches[:] # copy it
        for p in node_points: # linear scan
            if norm(new_point,p) < radius:
                knn_matches_out[1].append([list(p)])
                if len(knn_matches_out[1]) == k+1:
                    tmp = [norm(new_point,x) \
                           for x in knn_matches_out[1]]
                    del knn_matches_out[1][np.argmax(tmp)]
                    knn_matches_out[0] = np.min(tmp)
        else:
            dist_child_1 = norm(children[0][0],new_point)
            dist_child_2 = norm(children[1][0],new_point)
            node1 = None; node2 = None
            if dist_child_1 < dist_child_2:
                node1 = children[0]
                node2 = children[1]
            else:
                node1 = children[1]
                node2 = children[0]

```

```

knn_tmp = search_tree(new_point, knn_matches, node1, k)
knn_matches_out = search_tree(new_point, knn_tmp, node2, k)

return knn_matches_out

points = np.array([[3.,4.],[5.,5.],[9.,2.],[3.2,5.],[7.,5.],
                  [8.,9.],[7.,6.],[8,4],[6,2]])
tree = new_node()
form_tree(points,tree,all_points=points)
pp = pprint.PrettyPrinter(indent=4)
print "tree"
pp.pprint(tree)
newp = np.array([7.,7.])
dummysp = [np.Inf,np.Inf] # it should be removed immediately
res = search_tree(newp,[np.Inf, [dummysp]], tree, k=2)
print "done", res

tree
[ array([ 3.,  4.]),
  7.0710678118654755,
  None,
  [ [ array([ 8.,  9.]),
      3.1622776601683795,
      array([[ 8.,  9.],
             [ 7.,  6.]])],
    [None, None]],
    [ array([ 3.,  4.]),
      6.324555320336759,
      None,
      [ [ array([ 9.,  2.]),
          3.6055512754639891,
          None,
          [ [ array([ 7.,  5.]),
              1.4142135623730951,
              array([[ 7.,  5.],
                     [ 8.,  4.]])],
              [None, None]],
              [ array([ 9.,  2.]),
                3.0,
                array([[ 9.,  2.],
                       [ 6.,  2.]])],
              [None, None]]]],
      [ array([ 3.,  4.]),
        2.2360679774997898,
        None,
        [ [ array([ 5.,  5.]),
            0.0,
            array([[ 5.,  5.]])],
            [None, None]],
            [ array([ 3.,  4.]),
              1.019803902718557,
              array([[ 3.,  4.],
                     [ 3.2,  5. ]])],
            [None, None]]]]]]],
  None
done [1.0, [[[8.0, 9.0]], [[7.0, 6.0]]]]

```

Bu iki grup, o anda islemekte oldugumuz agac dugumun (node) iki cocuklari olacaktır. Çocuk noktaları kararlaştırıldıktan sonra artık sonraki asamaya gecilir, fonksiyon `form_tree` bu çocuk noktaları alarak, ayrı ayrı, her çocuk grubu için ozyineli (recursive) olarak kendi kendini çağırır. Kendi kendini çağırarak `form_tree`, tekrar başladığında kendini yeni (bir) nokta grubu ve yeni bir dugum objesi ile basbasa bulur, ve hiçbir şeyden habersiz olarak isleme koyulur. Tabii her ozyineli çağrı yeni dugum objesini yaratırken bir referansı üstteki ebeveyn dugume koymayı unutmamıştır, böylece ozyineli fonksiyon dünyadan habersiz olsa bile, ağacın en üstünden en altına kesintisiz bir bağlantı zinciri hep elimizde olur.

Not: `form_tree` içinde bir numara yaptık, tüm noktaların f_1 'e olan uzaklığı `dist1`, f_2 'e olan uzaklığı ise `dist2`. Sonra `diffs = dist1-dist2` ile bu iki uzaklığı birbirinden çıkartıyoruz ve mesela `points[diffs <= 0]` ile f_1 'e yakın olanları buluyoruz, çünkü bir tarafta f_1 'e yakınlık 4 diğer tarafta f_2 'ye yakınlık 6 ise, $4-6=-2$ ie o nokta f_1 'e yakın demektir. Ufak bir numara ile Numpy dilimleme (slicing) teknigini kullanabilmiş olduk ve bu önemli çünkü böylece `for` dongusu yazmıyoruz, Numpy'in arka planda C ile yazılmış hızlı rutinlerini kullanıyoruz.

Ek bazı bilgiler: kurelerin sınırları kesisebilir.

Arama

Üstte sözde program (pseudocode) BallKNN olarak gösterilen ve bizim kodda `search_tree` olarak anılan fonksiyon arama fonksiyonu. Aranılan `new_point`'e olan k en yakın diğer veri noktaları. Disaridan verilen değişken `knn_matches` üzerinde fonksiyon ozyineli bir şekilde arama yaparken "o ana kadar bulunmuş en yakın k nokta" ve o noktaların `new_point`'e olan en yakın mesafesi saklanır, arama işleyisi sırasında `knn_matches`, `knn_matches_out` sürekli verilip geri döndürülen değişkenlerdir, sözde programdaki P^{in} , P^{out} 'un karşılığidirler.

Arama algoritması şöyle işler: şimdi önceden oluşturulmuş kure hiyerarisisini üstten alta doğru gezmeye başlarız. Her basamakta yeni nokta ile o kurenin mi-henkini, yarıcapını kullanarak bir "alt sınır mesafe hesabı" yaparız, bu mesafe hesabının arkasında yatan düşünceyi yazının başında anlatmıştık. Bu mesafe kure içindeki tüm noktalara olan bir en az mesafe idi, ve eğer eldeki `knn_matches` üzerindeki şimdiye kadar bulunmuş mesafelerin en azından daha az ise, o zaman bu kure "bakmaya değer" bir kuredir, ve arama algoritması bu kureden işleme devam eder. Şimdiye kadar bulunmuş mesafelerin en azı `knn_matches` veri yapısı içine `min_so_far` olarak saklanıyor, sözde programdaki D_{sofar} .

Bu irdeleme sonrası (yani vs kuresinden yola devam kararı arkasından) işleme iki şekilde devam edilebilir, çünkü bir kure iki türden olabilir; ya nihai en alt kurelerden biridir ve üzerinde gerçek noktalar depolanmıştır, ya da ara kurelerden biridir (sona gelmedik ama doğru yoldayız, daha alta inmeye devam), o zaman fonksiyon yine ozyineli bir şekilde bu kurenin çocuklarına bakacaktır - her çocuk için kendi kendini çağıracaktır. İkinci durumda, kurede noktalar depolanmıştır, artık basit lineer bir şekilde o tüm noktalara teker teker bakılır, eldekilerden daha yakın olanı alınır, eldeki liste sısmeye başlamışsa (k 'den daha fazla ise) en büyük

noktalardan biri atilir [3], vs.

Daha alta inmemiz gereken birinci durumda yapilan iki cagrinin bir ozelligine dikkat cekmek isterim. Yeni noktanin bu cocuklara olan uzakligi da olculuyor, ve en once, en yakin olan cocuga dogru bir ozyineleme yapiliyor. Bu nokta cok onemli: niye boyle yapildi? Cunku icinde muhtemelen daha yakin noktalarin olabilecegi kurelere dogru gidersek, ozyineli cagrilarin teker teker bitip yukari dogru cikmaya baslamasi ve kaldiklari yerden bu sefer ikinci cocuk cagrilarini yapmaya baslamasi ardindan, elimizdeki `knn_matches` uzerinde en yakin noktalar buyuk bir ihtimalle zaten bulmus olacagiz. Bu durumda ikinci cagri yapilsa bile tek bir alt sinir hesabi o kurede dikkate deger hicbir nokta olamayacagini ortaya cikaracak (cunku en iyiler zaten elimizde), ve ikinci cocuga olan cagrilar hic alta inmeden pat diye geri donecektir, hic asagi inilmeyecektir.

Bu muthis bir kazanimdir: zaten bu stratejiye liteturde "budamak (pruning)" adi veriliyor, bu da cok uygun bir kelime aslinda, cunku agaclarla ugrasiyoruz ve bir dugum (kure) ve onun altindaki hicbir alt kureye ugramaktan kurtularak o dallarin tamamini bir nevi "budamis" oluyoruz. Bir suru gereksiz islemden de kurtuluyoruz bu arada, ve aramayi hizlandiriyoruz.

Mesafeler

Algoritmanin mesafeleri anlatan kisminde norm ve uzaylar gibi kavramlardan bahsettik. Yeni noktanin mihenke olan uzakliginin o kure icindeki tum diger noktalara olan uzakligini temsil edebilecegini soyledik: peki niye bu kavramlari direk bu sekilde anlatmadik, ve norm, ucgensel esitsizlik gibi kavramlardan bahsettik? Cunku 2 ve 3 boyut sonrasi uzaylari gorsel olarak dusunmek mumkun degildir, istedigimiz kadar ellerimizi kollarimizi sallayalim, bu kavramlari gorsel olarak tarif edemeyiz, ve degisik bir norm (mesafe) olcutu kullanmayi secebiliriz. Bu her iki durumda da elimizde soyut matematik baglaminda saglam bir temel oldugunu bilmek algoritmanin genelligini, ve degisik sartlarda uygulanabilirliğini arttirir. Mesela Oklit mesafesi yerine Manhattan mesafesi kullansam bile, bu mesafenin olcutunun norm kurallarini uydugunu bildigim icin kNN yapisinin geri kalanini oldugu gibi kullanabilirim, cunku o yapinin gecerlilikini normlar uzerinde gecerli ucgensel esitsizlik uzerinde ispat ettim.

Model

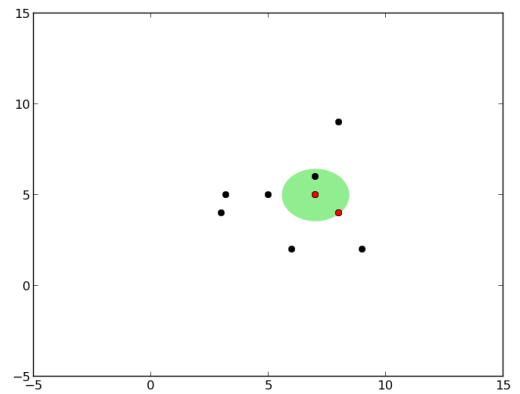
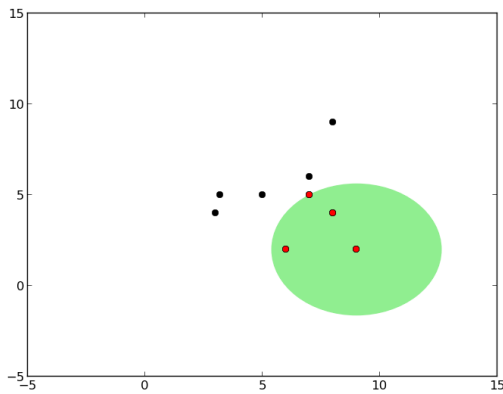
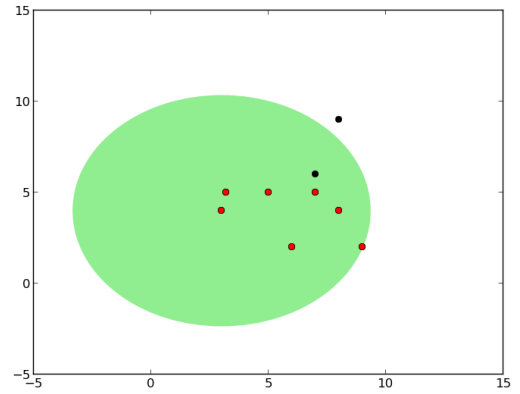
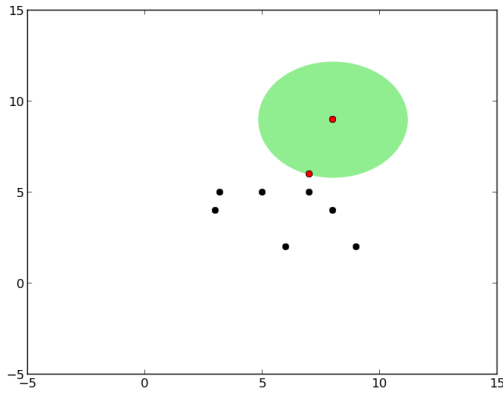
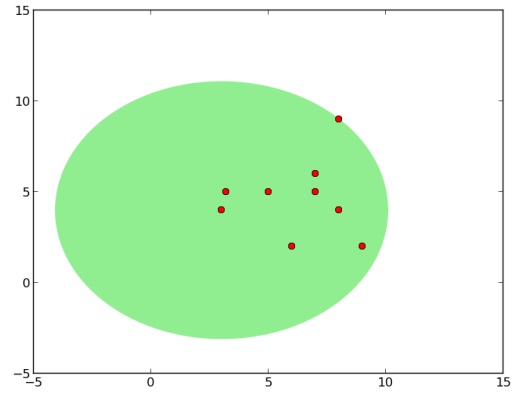
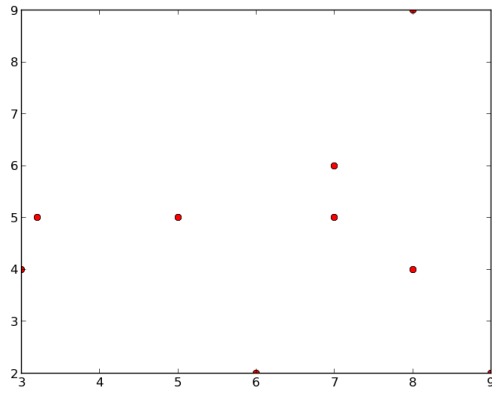
kNN'in model kullanmayan, model yerine verinin kendisini kullanan bir algoritma olarak tanittik. Peki "egitim" evresi sonrasi ele gecen kureler ve agac yapisi bir nevi model olarak gorulebilir mi?

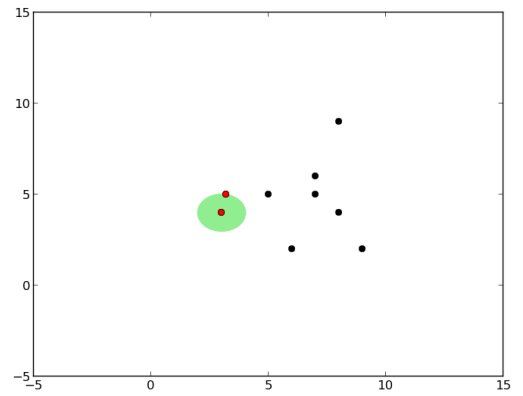
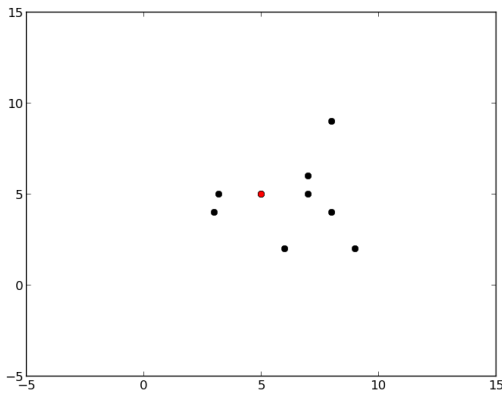
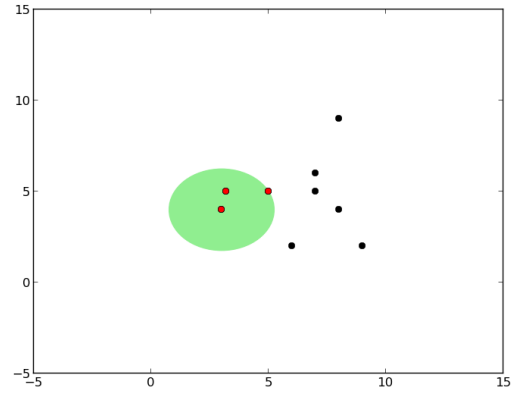
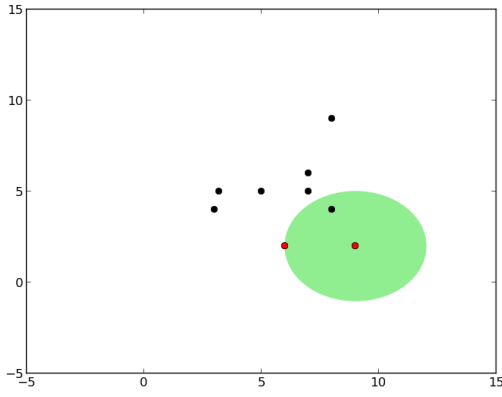
Bu onemli bir soru, ve bir bakima, evet agac yapisi sanki bir modelmis gibi duruyor. Fakat, mesela istatistiksel, grafiksel, yapay sinir aglari (neural net) baglaminda bakilrsa bu yapiya tam bir model denemez. Model bazli metotlarda model kurulunca veri atilir, ona bir daha bakilmaz. Fakat kNN, kure ve agac yapisini hala eldeki veriye erismek icin kullanmaktadir. Yani bir bakima veriyi "indeksliyoruz", ona erisimi kolaylastirip hizlandiriyoruz, ama ondan model cikartmiyoruz.

Not: Verilen Python kodu ve algoritma yakin noktaları hesaplıyor sadece, onların etiketlerinden hareketle yeni noktanın etiketini tahmin etme aşamasını gerçekleştiriyor. Fakat bu son aşama isin en basit tarafı, eğitim veri yapısına eklenecek bir etiket bilgisi ve sınıflama sonrası k noktanın ağırlıklı etiketinin hesabı ile basit şekilde gerçekleştirilebilir.

!python plot_circles.py

Agaci olusumu sirasinda kurelerin grafigi alttadir.





Kaynaklar, Notlar

[1] Liu, Moore, Gray, New Algorithms for Efficient High Dimensional Non-parametric Classification

[2] Alpaydın, Introduction to Machine Learning

[3] Silme islemi ornek kodumuzda Python `del` ile gercekleştirildi. Eger bu islem de hizlandirilmak istenirse, en alt kure seviyesindeki veriler bir oncelik kuyrug (priority queue) üzerinde tutulabilir, ve silme islemi hep en sondaki elemani siler, ekleme islemi ise yeni elemani (hep sirali olan) listede dogru yere koyar.