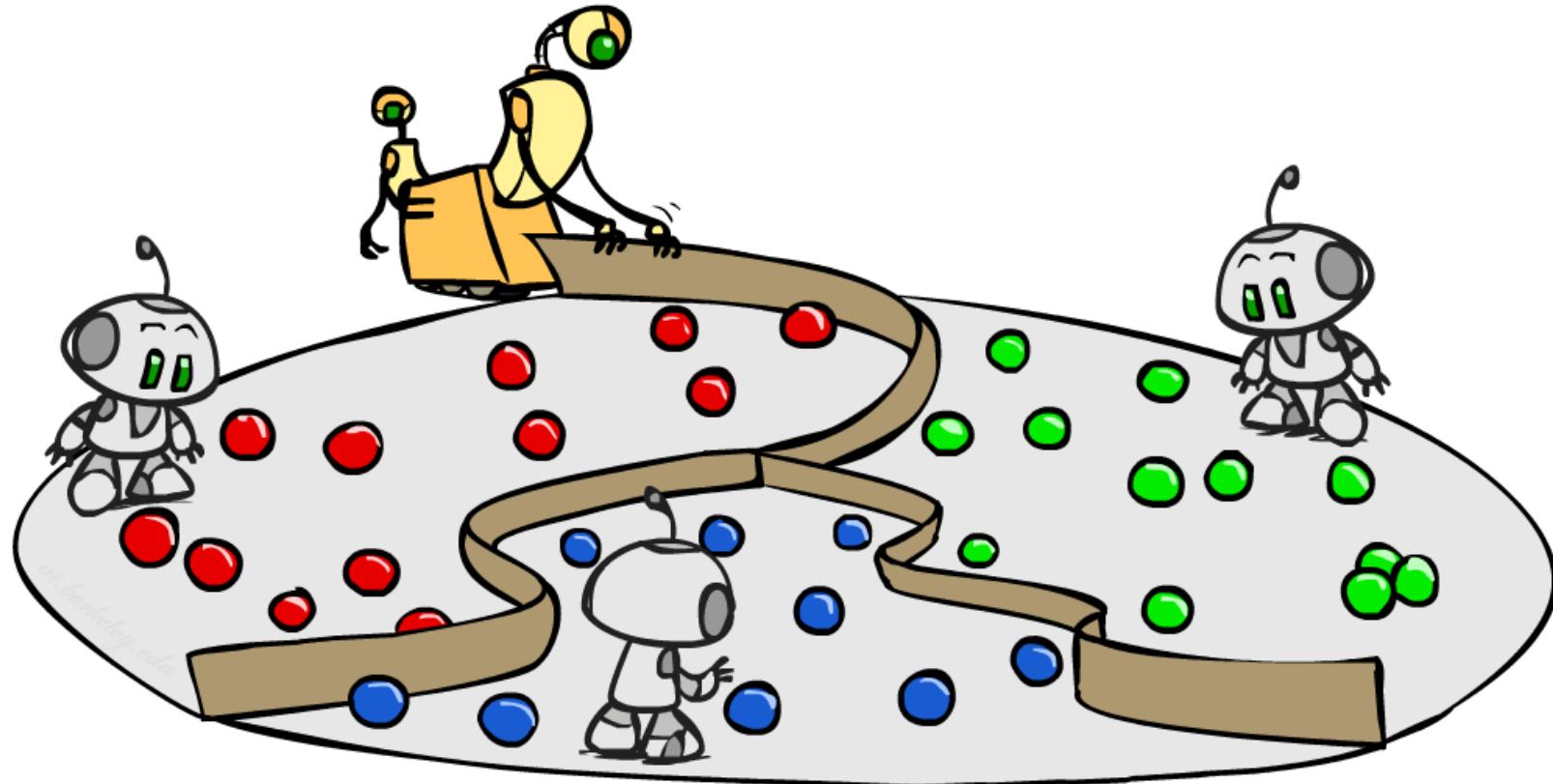


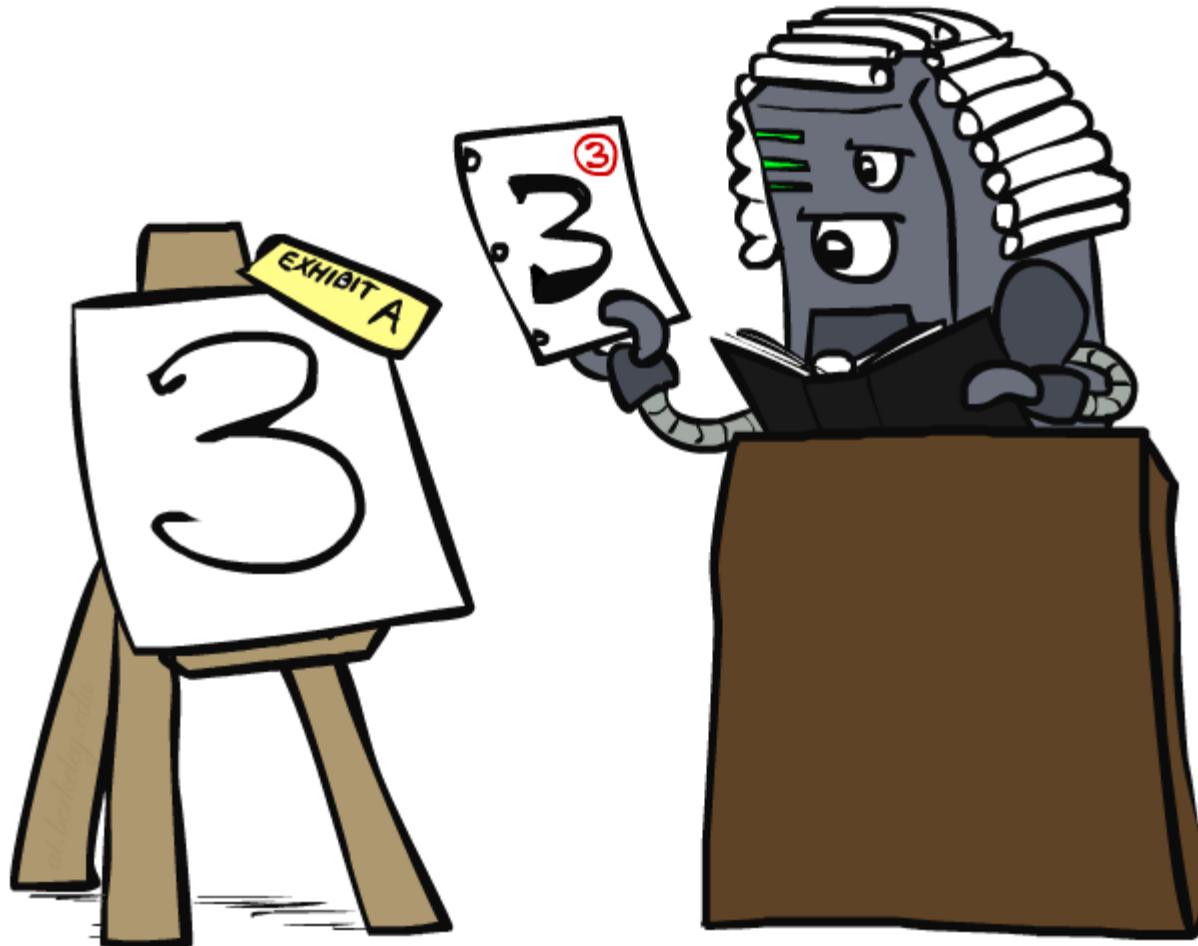
CS 188: Artificial Intelligence

Kernels and Clustering

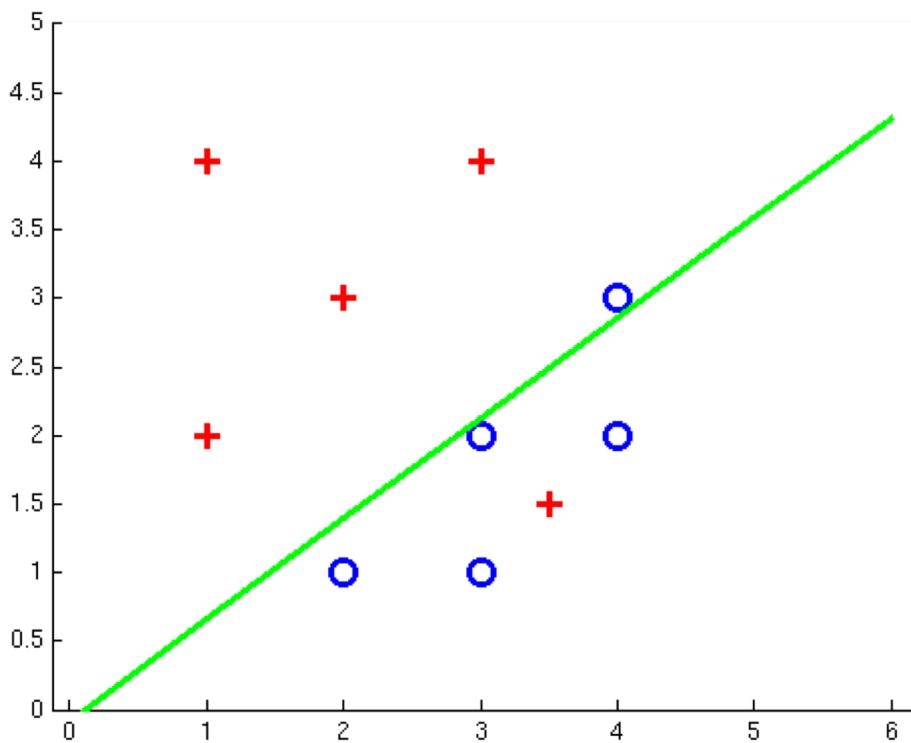


[These slides were created by Dan Klein and Pieter Abbeel and are available at <http://ai.berkeley.edu>.]

Case-Based Learning



Non-Separable Data



Case-Based Reasoning

به جای اینکه یک سری پارامتر برای جدا کردن تمام داده ها محاسبه بکند
داده جدید که رسیده را با داده های قبلی مقایسه می کند و اگر داده ای را پیدا کرد
بسیار شبیه به داده جدید است، داده جدید را در آن کلاس قرار می دهیم

Classification from similarity

- Case-based reasoning
- Predict an instance's label using similar instances

Nearest-neighbor classification

- 1-NN: copy the label of the most similar data point
- K-NN: vote the k nearest neighbors (need a weighting scheme)
- Key issue: how to define similarity
- Trade-offs: Small k gives relevant neighbors, Large k gives smoother functions

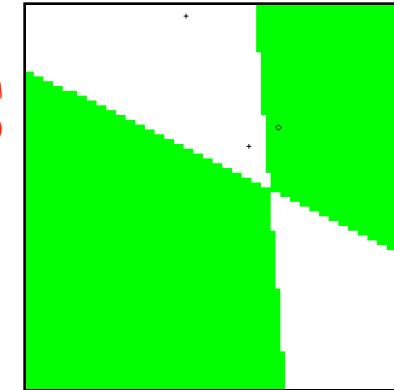


Parametric / Non-Parametric

■ Parametric models:

- Fixed set of parameters
- More data means better settings

در این روش اگر داده کمی داشته باشیم احتمال خطای آن بسیار بالا می رود و هر چه داده بیشتر شود دقیق این روش بالاتر میرود.



Truth

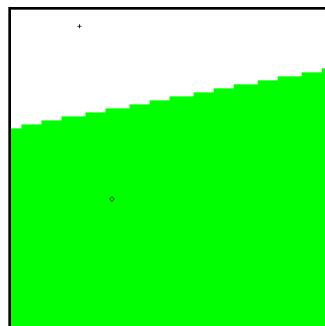
■ Non-parametric models:

- Complexity of the classifier increases with data
- Better in the limit, often worse in the non-limit

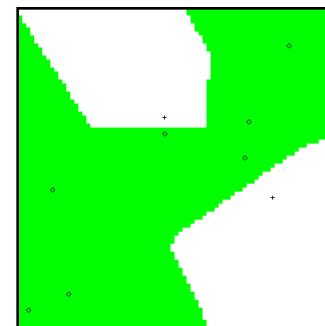
■ (K)NN is non-parametric

در شکل زیر مربوط برابر است با عمود منصف خط بین دو نقطه

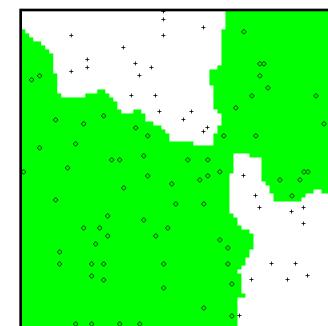
2 Examples



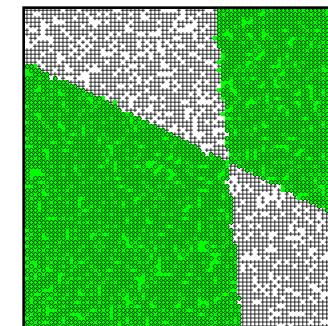
10 Examples



100 Examples



10000 Examples



Nearest-Neighbor Classification

- Nearest neighbor for digits:

- Take new image
- Compare to all training images
- Assign based on closest example



این ضرب داخلی جایی صفر می شود که این دو داده هیچ اشتراکی با هم نداشته باشند

این ضرب داخلی زمان یک می شود که دو داده دقیقاً یک تصویر باشد (هر چیز را با خودش مقایسه کنیم یک می شود)

- Encoding: image is vector of intensities:

$$\begin{matrix} 1 \\ \text{---} \end{matrix} = \langle 0.0 \ 0.0 \ 0.3 \ 0.8 \ 0.7 \ 0.1 \dots 0.0 \rangle$$



- What's the similarity function?

- Dot product of two images vectors?

برای هر داده متوسط همه رو به دست بیاریم و از هر نقطه آن را کم کنیم
که یک بردار نرمال میشود و با این روش روی صفر آن را نرمال کردیم



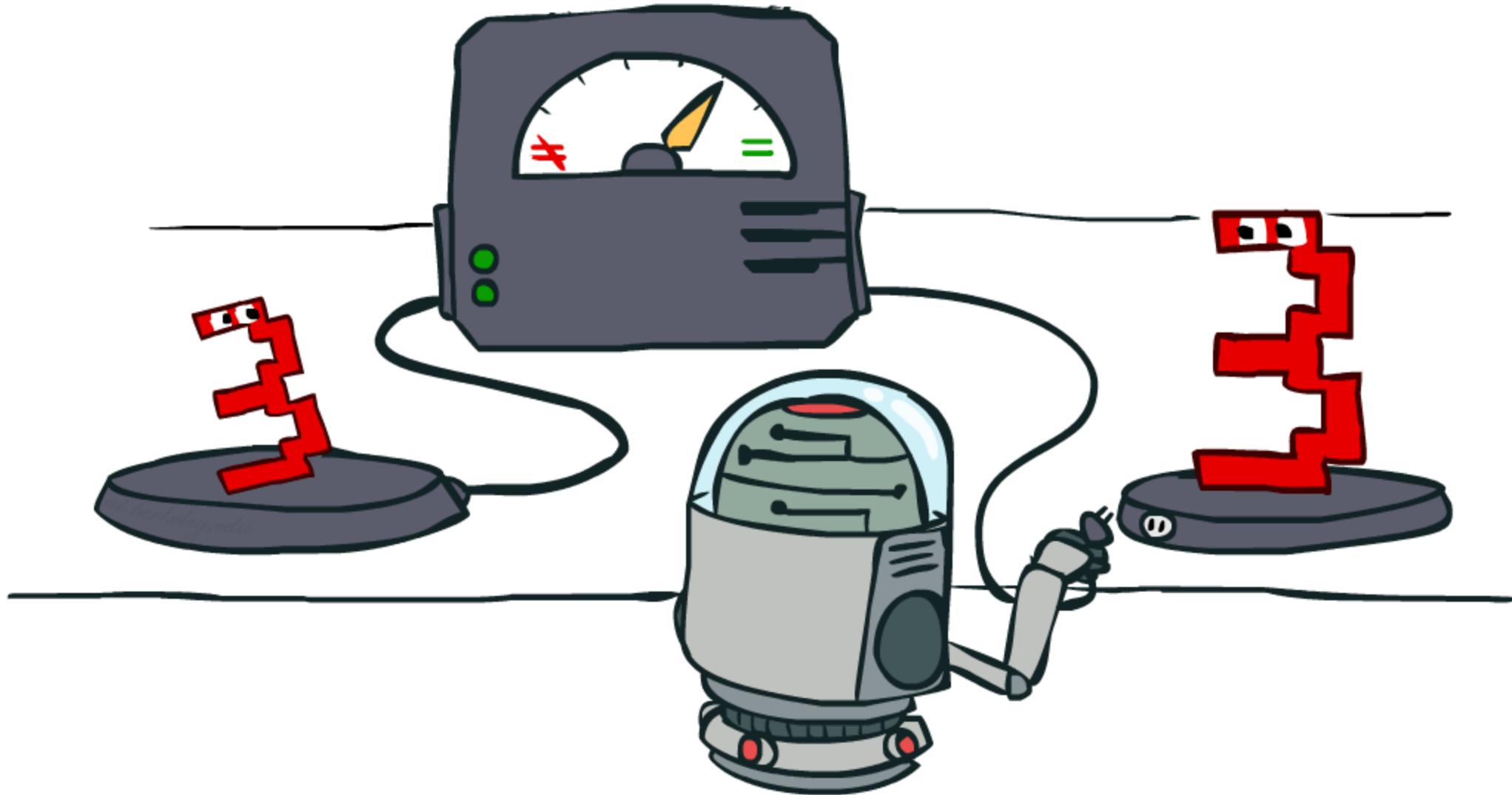
$$\text{sim}(x, x') = x \cdot x' = \sum_i x_i x'_i$$



- Usually normalize vectors so $\|x\| = 1$
- $\min = 0$ (when?), $\max = 1$ (when?)



Similarity Functions



Basic Similarity

- Many similarities based on **feature dot products**:

$$\text{sim}(x, x') = f(x) \cdot f(x') = \sum_i f_i(x)f_i(x')$$

- If features are just the pixels:

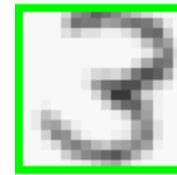
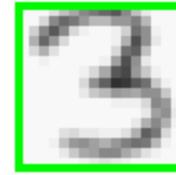
$$\text{sim}(x, x') = x \cdot x' = \sum_i x_i x'_i$$

- Note: not all similarities are of this form

Invariant Metrics

- Better similarity functions use knowledge about vision
- Example: invariant metrics:

- Similarities are invariant under certain transformations
- Rotation, scaling, translation, stroke-thickness...
- E.g:



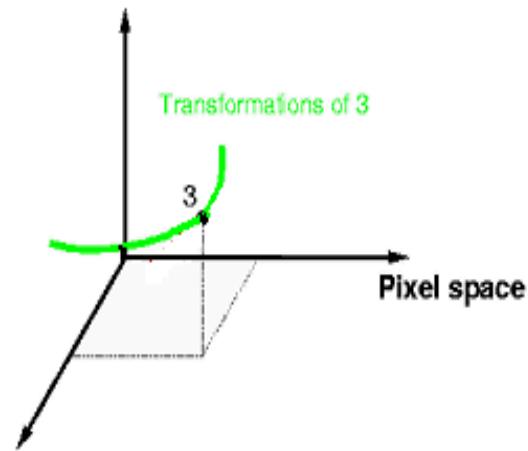
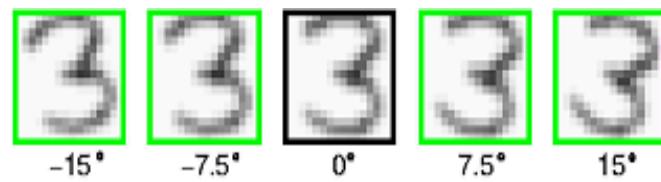
اگر معیار مقایسه درستی داشته باشیم نباید چرخش و یا هر تغییر دیگری آن را از مقایسه درست منحرف کند

در این مثال با توجه به معیاری که ما گفتیم شباهت این دو داده بسیار کم است

برای این منظور عواملی که باعث کم شدن شباهت می شود را هم لحاظ می کنند
یعنی به جای اینکه وقتی یک داده جدید میاد فقط با اون داده train مقایسه بکنند
با یک مجموعه داده های شبیه داده های train مقایسه می کنند
مثالاً: داده train را ۵ درجه می چرخانیم و هی مقایسه می کنیم و هر جا مаксیمم شد
آن را در نظر می گیریم
و در واقع به این معناست که هر داده train را کمی دوران داده و با همان برچسب دوباره داخل داده های train اضافه می کنیم
و بعد معیار شباهت را می سنجیم

- $16 \times 16 = 256$ pixels; a point in 256-dim space
- These points have small similarity in R^{256} (why?)
- How can we incorporate such invariances into our similarities?

Rotation Invariant Metrics



- Each example is now a curve in \mathbb{R}^{256}
- Rotation invariant similarity:

$$s' = \max s(r(\boxed{\text{3}}), r(\boxed{\text{3}}))$$

- E.g. highest similarity between images' rotation lines

Template Deformation

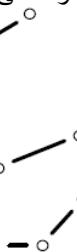
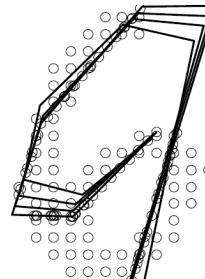
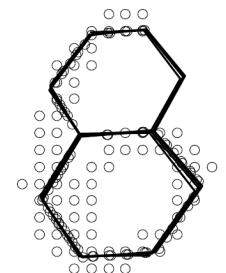
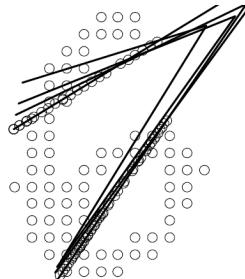
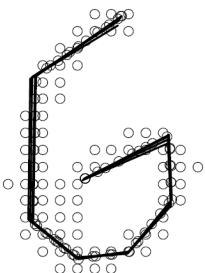
روش ۳

■ Deformable templates:

- An “ideal” version of each category
- Best-fit to image using min variance
- Cost for high distortion of template
- Cost for image points being far from distorted template

■ Used in many commercial digit recognizers

تغییری را که میدهیم نباید و بسته به کلاس باشد
مثالاً ۶ را نباید آن قدر چرخاند که ۹ شود و این را
باید در هنگام تغییرات لحاظ کرد



می بوان

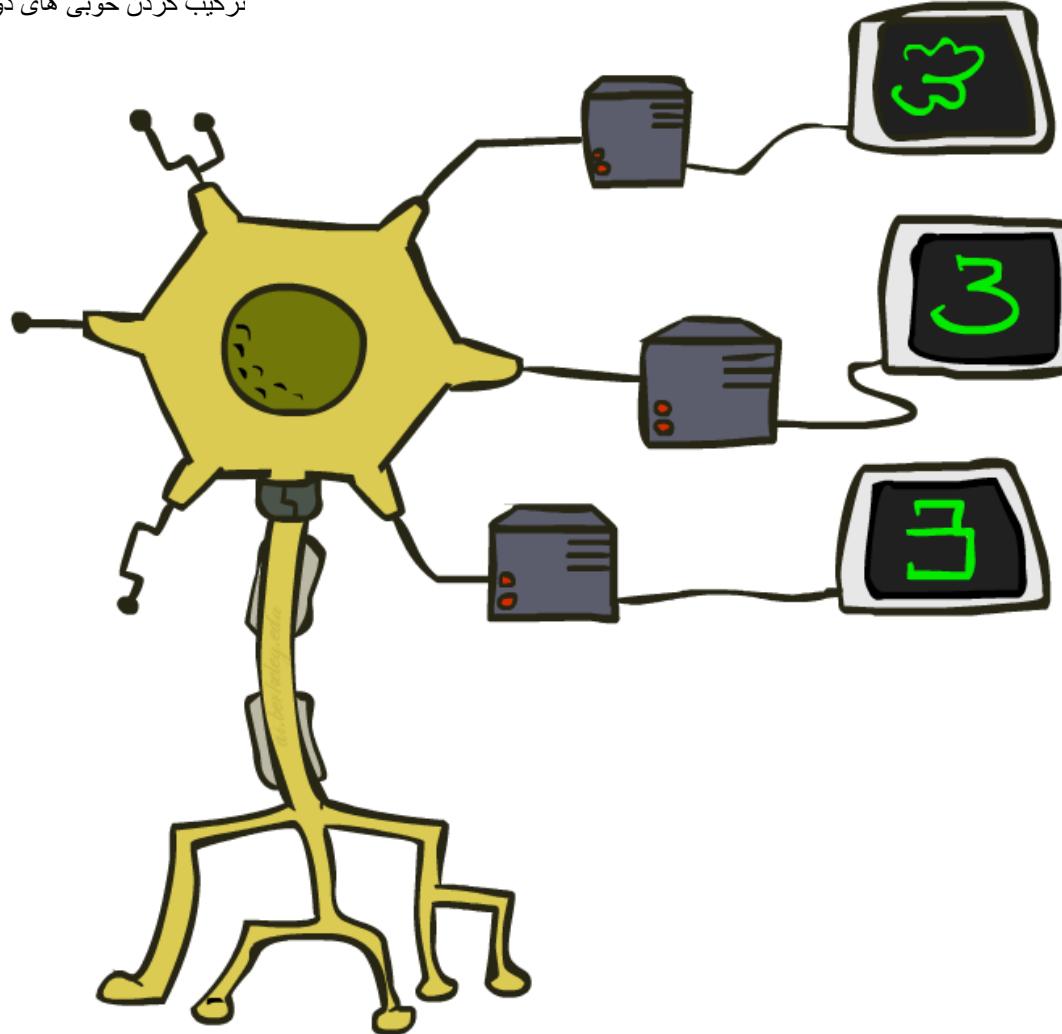
به بهترین شکل تبدیل کرد
و به جای این که فاصله را حساب کنیم این هزینه را حساب می کنیم و داده ای که
منimum هذنه، ۱ دارد

A Tale of Two Approaches...

- Nearest neighbor-like approaches
 - Can use fancy similarity functions
 - Don't actually get to do explicit learning
- Perceptron-like approaches
 - Explicit training to reduce empirical error
 - Can't use fancy similarity, only linear
 - Or can they? Let's find out!

Kernelization

ترکیب کردن خوبی های دو روش perceptron و k nearest neighbour



Perceptron Weights

- What is the final value of a weight w_y of a perceptron?

- Can it be any real vector?
- No! It's built by adding up inputs.

بردار وزن ها W_y نمی تواند هر بردار باشد
 بلکه تنها باید ترکیب خطی بردار های داده ها باشد

در این مثال x_1 اشتباه دسته بنده شده (باید مثبت می داده ولی منفی داده) پس باید مثبت ترش کنیم بنابراین بردار وزن ها را با

بردار $(f(x_1), f(x_2), f(x_3), f(x_4))$ جمع می کنیم

$f(x_2), f(x_3), f(x_4) \rightarrow \text{correct}$

بردار x_5 اشتباه تشخیص داده می شده (باید منفی می بوده ولی مثبت تشخیص دادیم) بنابراین $f(x_5)$ رو از بردار وزن ها کم می کنیم

و

$$w_y = 0 + f(x_1) - f(x_5) + \dots$$

$$w_y = \sum_i \alpha_{i,y} f(x_i)$$

ALPHA_{i,y}
تأثیر داده آم در بردار کلاس y

به جای update کردن بردار وزن ها بردار y را ALPHA_{i,y} update کنیم

یک مشکل واضح :

تعداد چیز هایی که نگه می داریم ممکن است خیلی زیاد شود

قبل از تعداد feature ها عدد نگه می داشتیم

در حالی که الان تعداد پارامتر ها بیشتر شده است

- Can reconstruct weight vectors (the primal representation) from update counts (the dual representation)

$$\alpha_y = \langle \alpha_{1,y} \ \alpha_{2,y} \ \dots \ \alpha_{n,y} \rangle$$

Dual Perceptron

- How to classify a new example x ?

همان کارهایی که با بردار w می کردیم را حالا باید با بردا آلفا انجام دهیم :

$$\begin{aligned}\text{score}(y, x) &= w_y \cdot f(x) \\ &= \left(\sum_i \alpha_{i,y} f(x_i) \right) \cdot f(x) \\ &= \sum_i \alpha_{i,y} (f(x_i) \cdot f(x)) \\ &= \sum_i \alpha_{i,y} K(x_i, x)\end{aligned}$$

از روی بردار آلفا و بردار داده x (داده جدید) که
داریم و x_i تمام داده های train می باشد

- If someone tells us the value of K for each pair of examples, never need to build the weight vectors (or the feature vectors)!

Dual Perceptron

- Start with zero counts (alpha)
- Pick up training instances one by one
- Try to classify x_n ,

$$y = \arg \max_y \sum_i \alpha_{i,y} K(x_i, x_n)$$

- If correct, no change!
- If wrong: lower count of wrong class (for this instance), raise count of right class (for this instance)

$$\alpha_{y,n} = \alpha_{y,n} - 1$$

$$w_y = w_y - f(x_n)$$

$$\alpha_{y^*,n} = \alpha_{y^*,n} + 1$$

$$w_{y^*} = w_{y^*} + f(x_n)$$

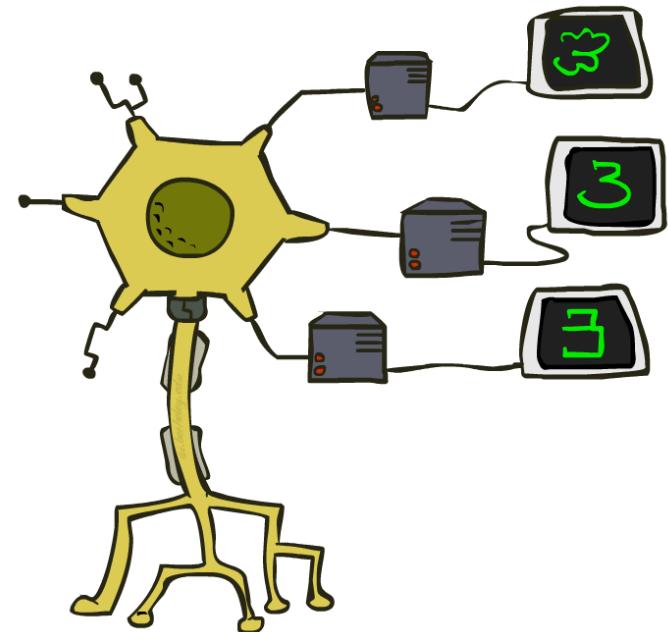
کلاسی که بیشتری score را می دهد انتخاب می کنیم
اگر درست بود که هیچی
اگر غلط بود بردار w را با بردار داده جمع یا تفریق می کردیم

Kernelized Perceptron

- If we had a black box (kernel) K that told us the dot product of two examples x and x' :
 - Could work entirely with the dual representation
 - No need to ever take dot products (“kernel trick”)

$$\begin{aligned}\text{score}(y, x) &= \mathbf{w}_y \cdot f(x) \\ &= \sum_i \alpha_{i,y} K(x_i, x)\end{aligned}$$

- Like nearest neighbor – work with black-box similarities
- Downside: slow if many examples get nonzero alpha



چه طور از یک معیار پیچیده شباهت استفاده کنیم؟
این k دو نقطه داده را می‌گیرد و میزان شباهت آن‌ها را با ضرب نقطه‌ای می‌سنجد
می‌توان این k را برداشت و هر چیزی به جای آن قرار دهیم

Kernels: Who Cares?

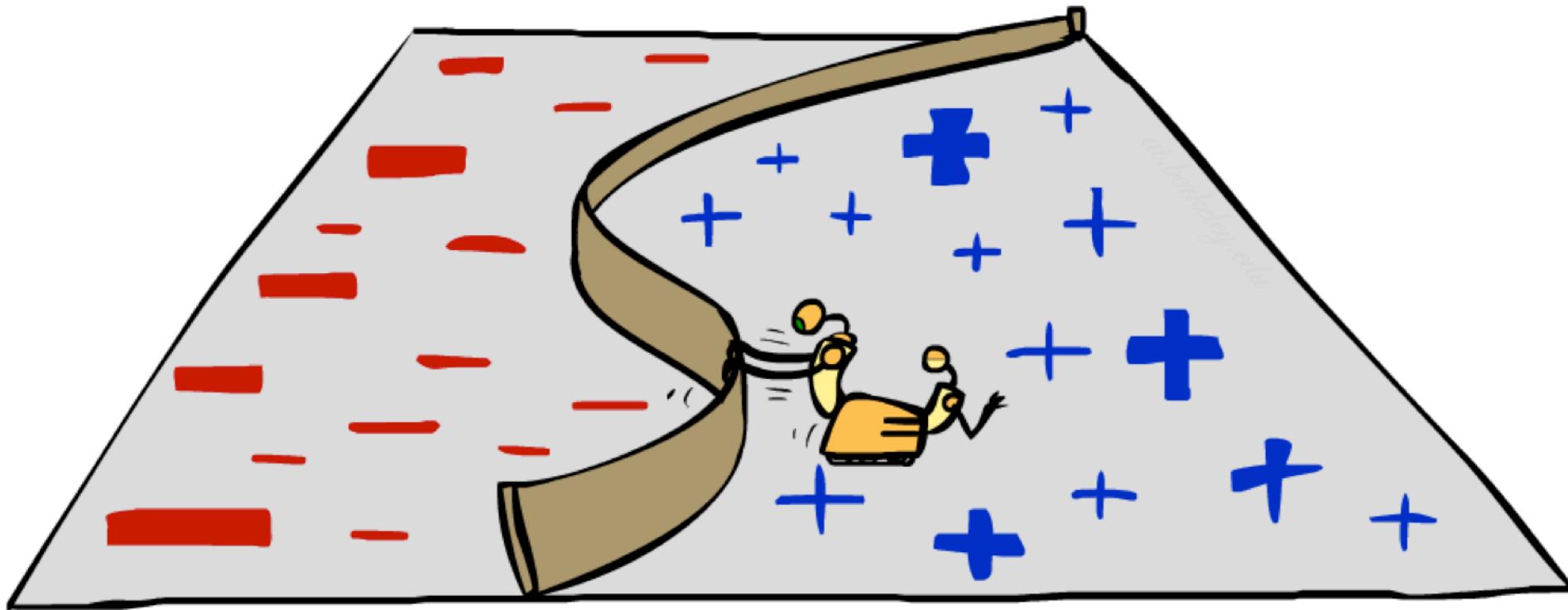
- So far: a very strange way of doing a very simple calculation
- “Kernel trick”: we can substitute any* similarity function in place of the dot product

این معیار شباهت می تواند این امکان را به ما بدهد که به جای خط های مستقیم بتوانیم خط های پیچیده تری برای جدا کردن کلاس ها در نظر بگیریم
اما چه طور ؟؟؟؟
- Lets us learn new kinds of hypotheses

* Fine print: if your kernel doesn't satisfy certain technical requirements, lots of proofs break. E.g. convergence, mistake bounds. In practice, illegal kernels sometimes work (but not always).

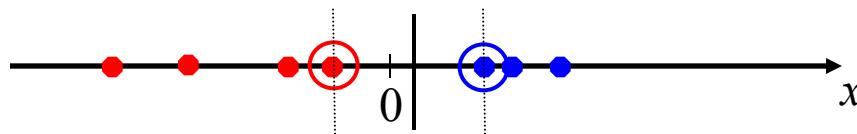
Non-Linearity

اگر داده ها خطی تفکیک باشند از همان استفاده می کنیم و اصلا نیازی نیست از perceptron استفاده کنیم kernel

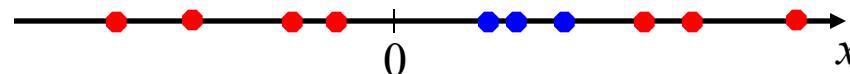


Non-Linear Separators

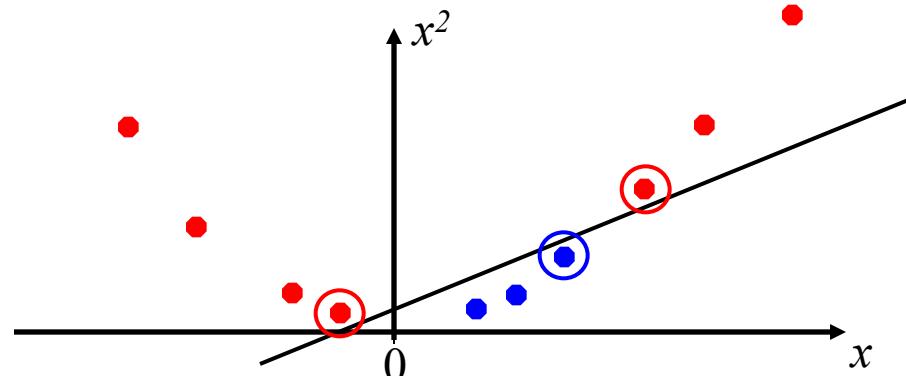
- Data that is linearly separable works out great for linear decision rules:



- But what are we going to do if the dataset is just too hard?



- How about... mapping data to a higher-dimensional space:

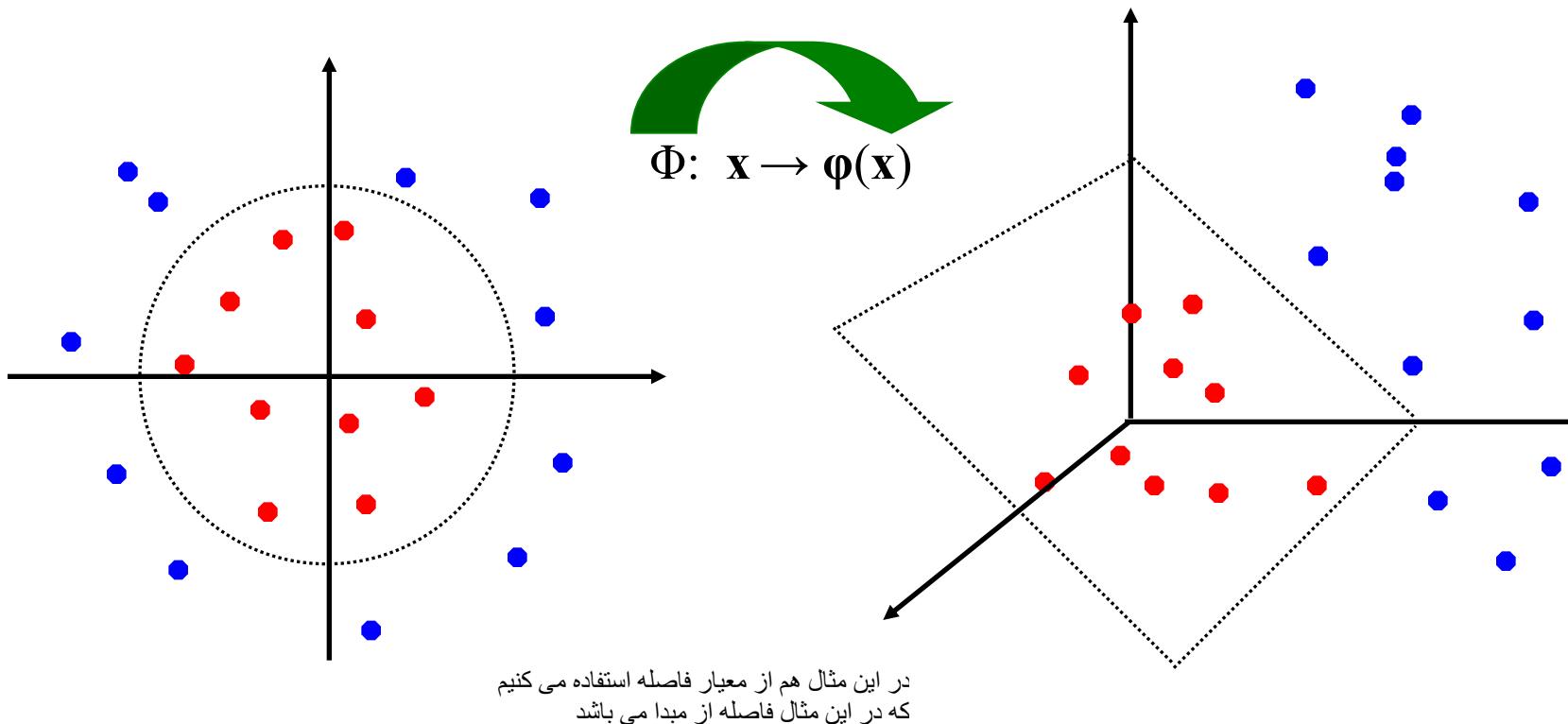


چرا x^2 یک feature مناسب می باشد؟
جنس معیاری که واقعا تاثیر گذارد است از جنس فاصله است نه مقدار
که فاصله همان x^2 می باشد

Non-Linear Separators

مثال ۲

- General idea: the original feature space can always be mapped to some higher-dimensional feature space where the training set is separable:



Some Kernels

- Kernels implicitly map original vectors to higher dimensional spaces, take the dot product there, and hand the result back

- Linear kernel:

$$K(x, x') = x' \cdot x' = \sum_i x_i x'_i$$

- Quadratic kernel:

$$K(x, x') = (x \cdot x' + 1)^2$$

از این کرنل‌ها استفاده کنیم در حالی که داده‌ها را به فضای بزرگتر می‌بریم ولی محاسبات ثابت می‌ماند

به همین دلیل به آن kernel trick می‌گویند !!
 

مثل این است که یک سری feature اضافه کردیم
و در فضای جدید از ضرب داخلی استفاده کردیم

$$= \sum_{i,j} x_i x_j x'_i x'_j + 2 \sum_i x_i x'_i + 1$$

- RBF: infinite dimensional representation

$$K(x, x') = \exp(-||x - x'||^2)$$

- Discrete kernels: e.g. string kernels

چه feature‌هایی اضافه کردیم ؟
یک بردار داریم x_1, x_2, x_3, x_4
انگار یک سری feature اضافه کردیم
 $x_1^2, x_1 x_2, x_1 x_3, x_1 x_4, x_2^2, \dots$
که به شکل بالا می‌باشد
که مثل به توان ۲ رساندن می‌باشد
بنابراین مرزهایی که می‌کشیم می‌تواند دایره
هذلولی یا دایره بکشیم

Why Kernels?

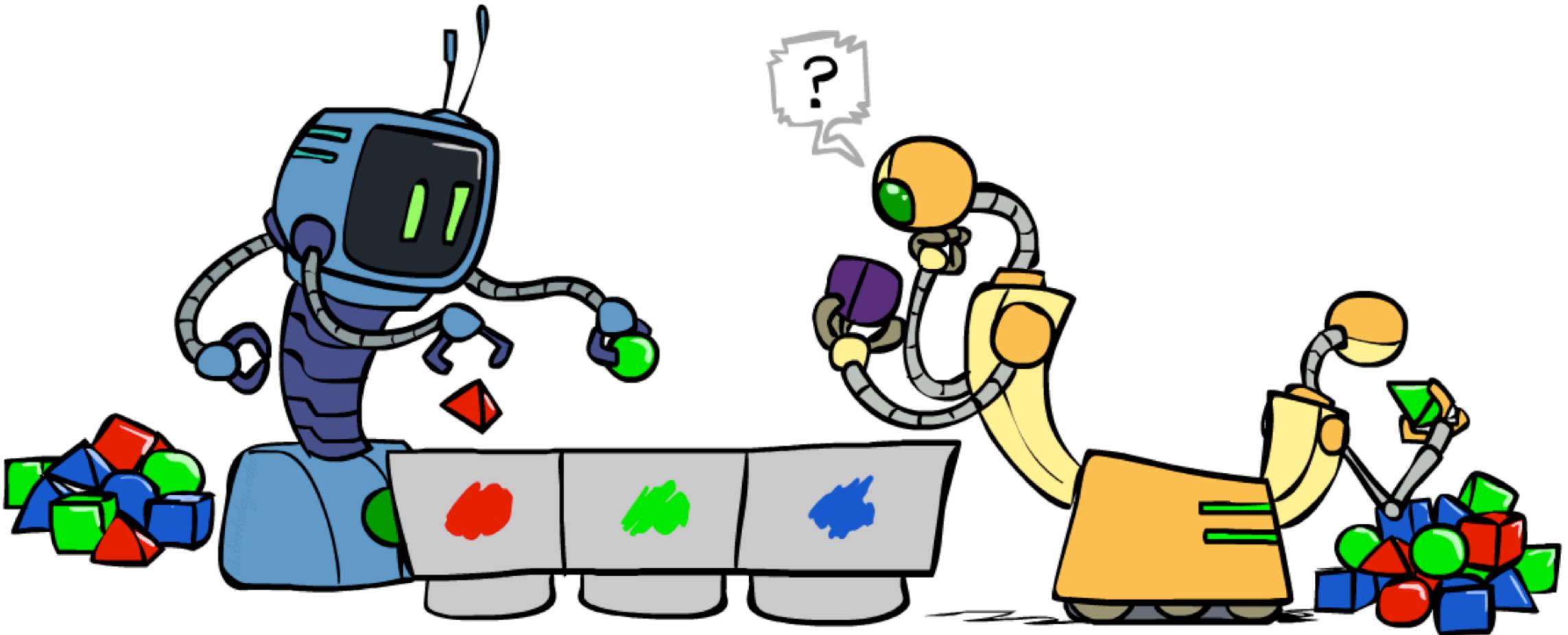
- Can't you just add these features on your own (e.g. add all pairs of features instead of using the quadratic kernel)?
 - Yes, in principle, just compute them
 - No need to modify any algorithms
 - But, number of features can get large (or infinite)
 - Some kernels not as usefully thought of in their expanded representation, e.g. RBF kernels
- Kernels let us compute with these features implicitly
 - Example: implicit dot product in quadratic kernel takes much less space and time per dot product
 - Of course, there's the cost for using the pure dual algorithms: you need to compute the similarity to every training datum

Clustering

- Clustering systems:
 - Unsupervised learning
 - Detect patterns in unlabeled data
 - E.g. group emails or search results
 - E.g. find categories of customers
 - E.g. detect anomalous program executions
 - Useful when don't know what you're looking for
 - Requires data, but no labels
 - Often get gibberish

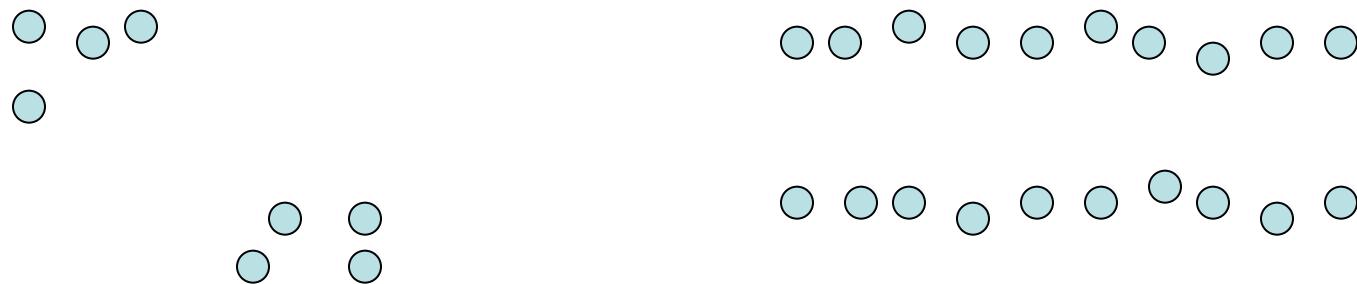


Clustering



Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns

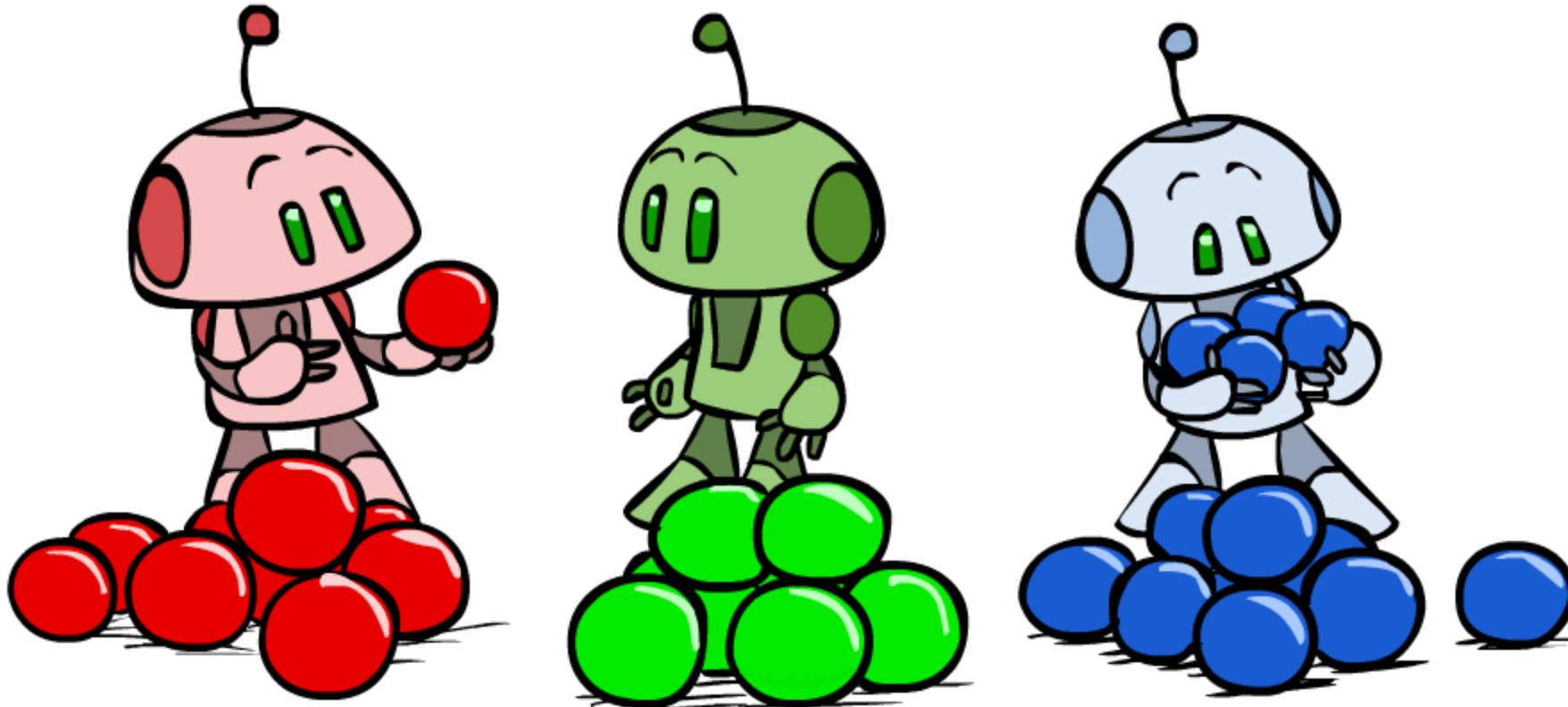


- برای این که برنامه ای بنویسیم که این کار را اتوماتیک انجام دهد
می توان از فاصله اقلیدسی برای معیار شباht استفاده کرد (ولی نه همیشه)
در واقع باید connectivity نقاط از طریق نقاط دیگر باید مهم باشد
- What could “similar” mean?
 - One option: small (squared) Euclidean distance

$$\text{dist}(x, y) = (x - y)^T (x - y) = \sum_i (x_i - y_i)^2$$

الگوریتم برای
clustering

K-Means



K-Means

- An iterative clustering algorithm

- Pick K random points as cluster centers (means)

- Alternate:

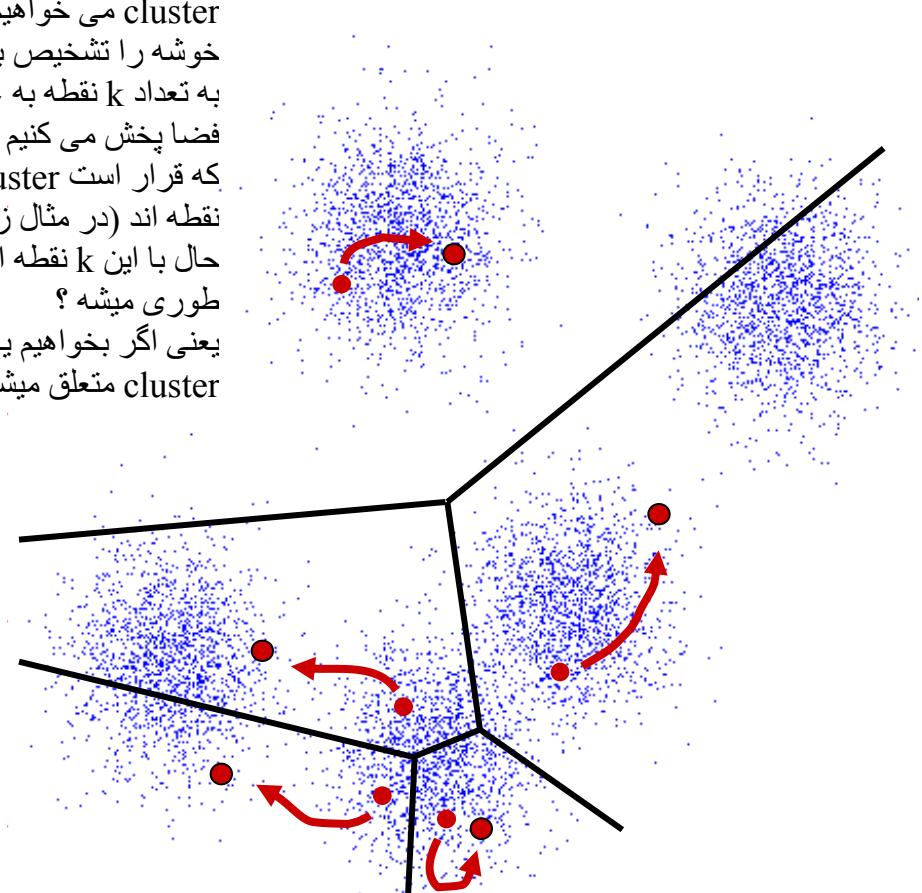
- Assign data instances to closest mean
- Assign each mean to the average of its assigned points

- Stop when no points' assignments change

چه طور این تعلق را انجام دهیم؟

اونی که به داده ورودی نزدیک تر است بنابراین داده جدید متعلق به این cluster است

در ابتدای کار الگوریتم مشخص می کنیم که چند cluster می خواهیم تشکیل دهیم. (می خواهیم k خوش را تشخیص بدهیم) به تعداد k نقطه به عنوان نماینده این خوش را در فضای پخش می کنیم در لحظه شروع می گوییم نقاطی که قرار است cluster ها را نمایندگی بکنند این k نقطه اند (در مثال زیر $k=5$) حال با این k نقطه ای گذاشتیم cluster ها چه طوری میشه؟ یعنی اگر بخواهیم یک داده جدید را ببینیم به کدام cluster متعلق میشه؟

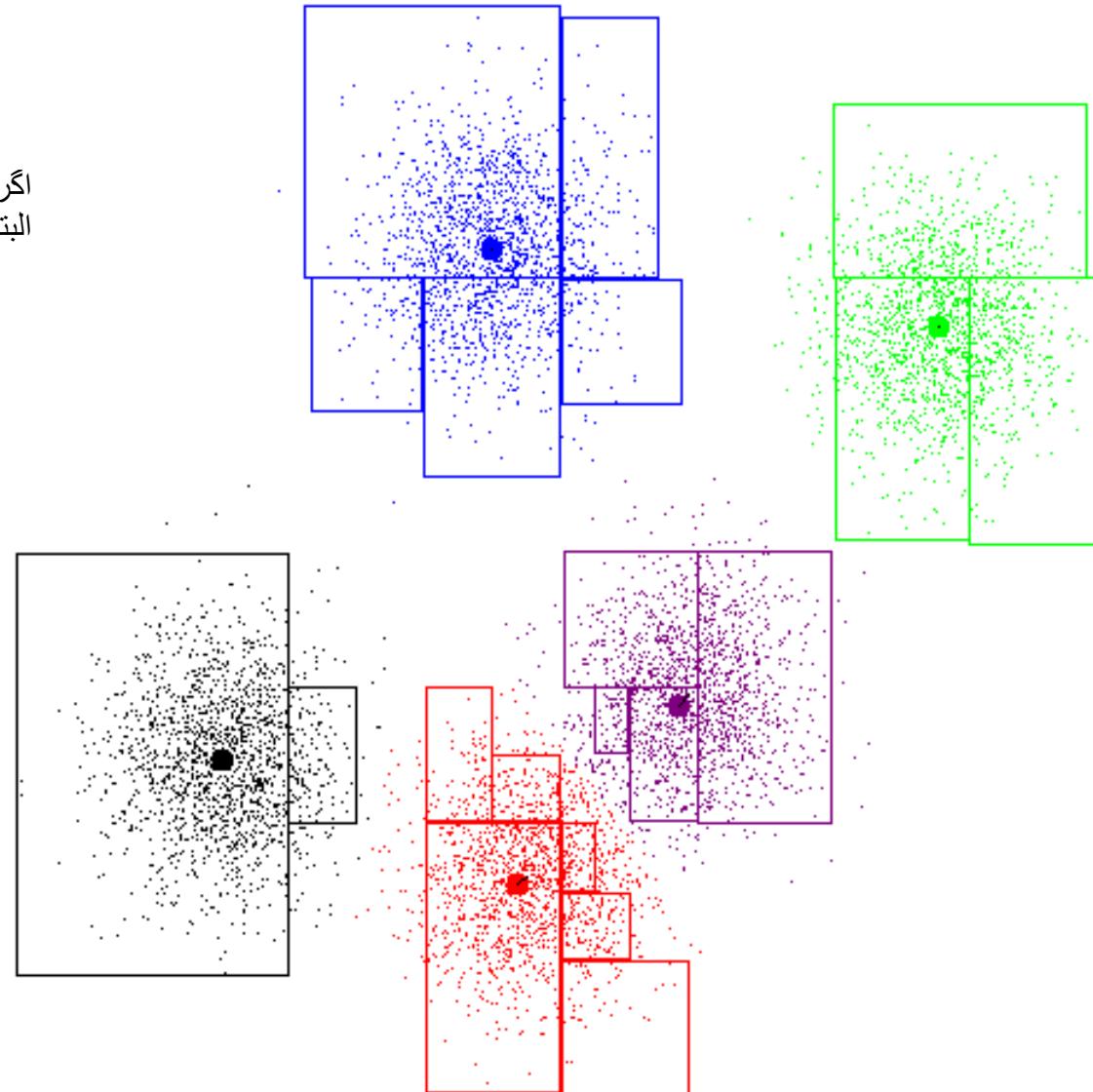


هر نقطه را به میانگین داده های آن کلاس منتقل می کنیم

مرزهایی که در شکل می بینیم هر کدام عمود منصف هر دو تایی از این نقاط است در واقع این نقاط قرار است در هر لحظه بیانگر میانگین نقاطی باشند که در مربوط به آن داده قرار دارند.

K-Means Example

اگر جای نقاط عوض بشه آیا جای مرزها نباید تغییر کند؟؟
البته که مرزها نیز باید جابجا شوند



K-Means as Optimization

- Consider the total distance to the means:

$$\phi(\{x_i\}, \{a_i\}, \{c_k\}) = \sum \text{dist}(x_i, c_{a_i})$$

points assignments means

- Each iteration reduces phi

- Two stages each iteration:

- Update assignments: fix means c, change assignments a
- Update means: fix assignments a, change means c

برای این که نشان دهیم با این تابع هزینه مثبت حتماً الگوریتم متوقف می‌شود:

یک راهش این است که نشان دهیم هر کاری که در الگوریتم انجام می‌دهیم این تابع را کم می‌کند

چه طور نشان دهیم این تابع در هر مرحله کم می‌شود؟

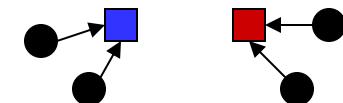
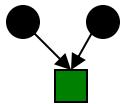
هر مرحله از الگوریتم: یا داریم تعلق نقاط را تغییر می‌دهیم (ولی زمانی این کار را می‌کنیم که جدید نزدیک‌تر باشد پس حالا یک عدد کوچکتری جمع می‌شود)

جای نماینده‌ها رابه میانگین نقاط تغییر می‌دهیم

آیا این الگوریتم متوقف می‌شود یا نه؟

چون هم جای نقاط را تغییر می‌دهیم و هم تعلق را عوض می‌کنیم، آیا ممکن است در دور بیافتیم؟

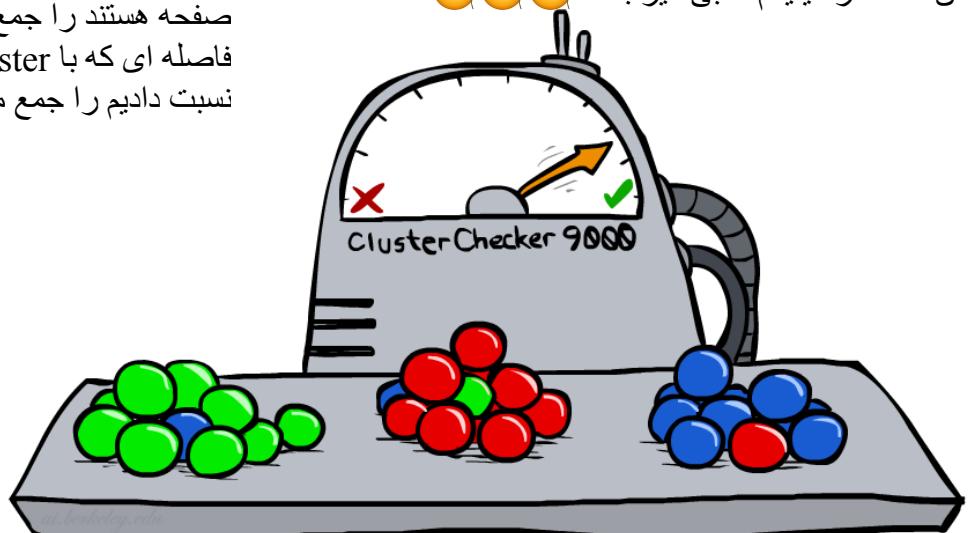
این الگوریتم قطعاً متوقف می‌شود، اما چه طور آن را ثابت می‌کنیم؟



آیا از این روش می‌توان بهترین k تا cluster را تعیین بکنیم؟

لزوماً نمی‌توان تضمین کرد تابع هزینه به \min مطلق خود برسد

ممکن است در مینیمم نسبی گیر بکند

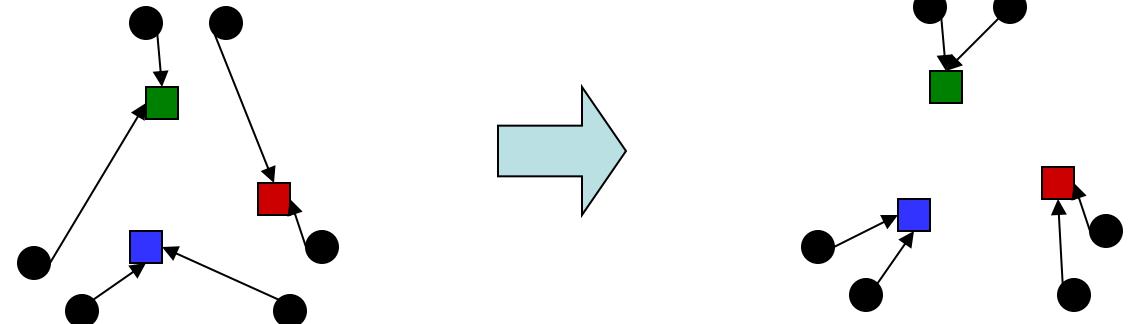


Phase I: Update Assignments

- For each point, re-assign to closest mean:

$$a_i = \operatorname{argmin}_k \text{dist}(x_i, c_k)$$

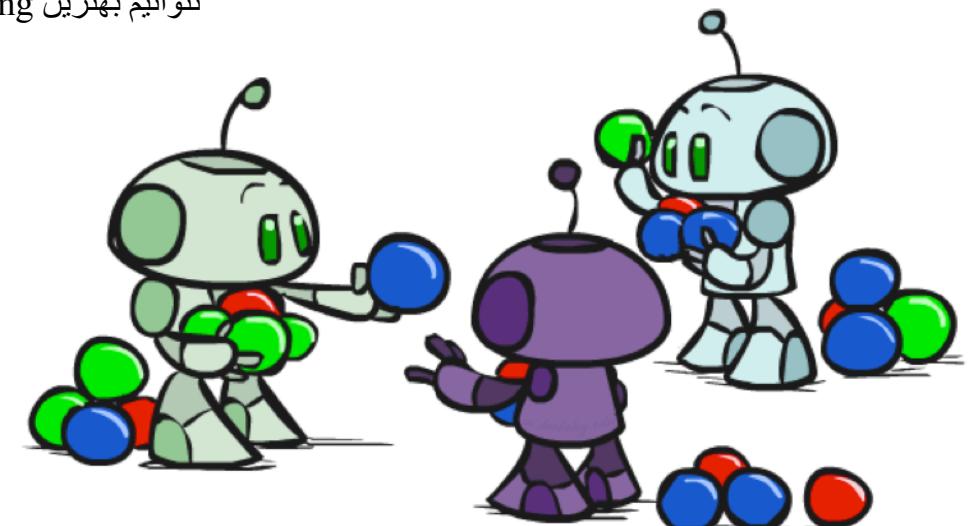
نکات در اسلاید ۳۰



- Can only decrease total distance phi!

$$\phi(\{x_i\}, \{a_i\}, \{c_k\}) = \sum_i \text{dist}(x_i, c_{a_i})$$

یک موضوع به تعداد k بر می گردد
ولی حتی وقتی به گفته شده چند خوش داشته باشیم
نتوانیم بهترین clustering را به دست آوریم

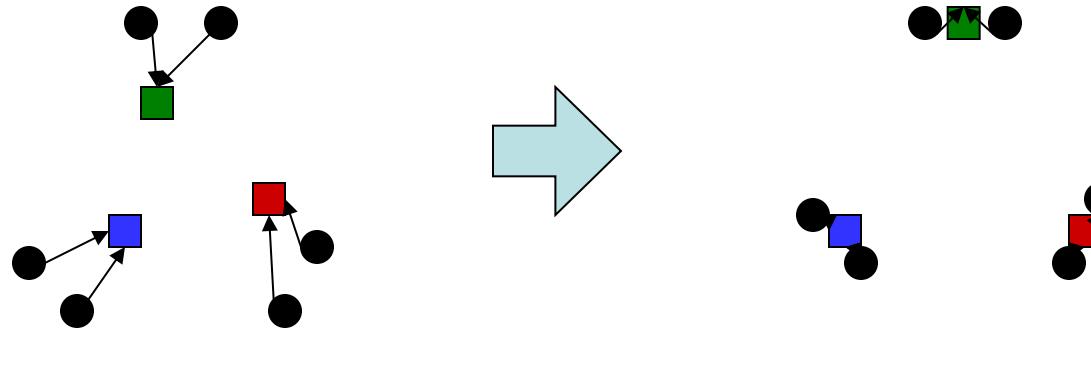


Phase II: Update Means

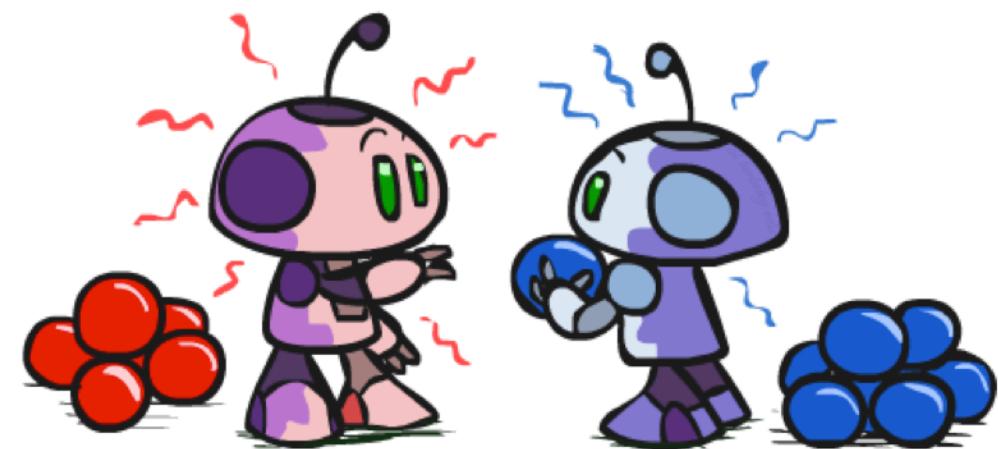
- Move each mean to the average of its assigned points:

$$c_k = \frac{1}{|\{i : a_i = k\}|} \sum_{i:a_i=k} x_i$$

٣٠ نکات در اسلاید



- Also can only decrease total distance... (Why?)
- Fun fact: the point y with minimum squared Euclidean distance to a set of points $\{x\}$ is their mean

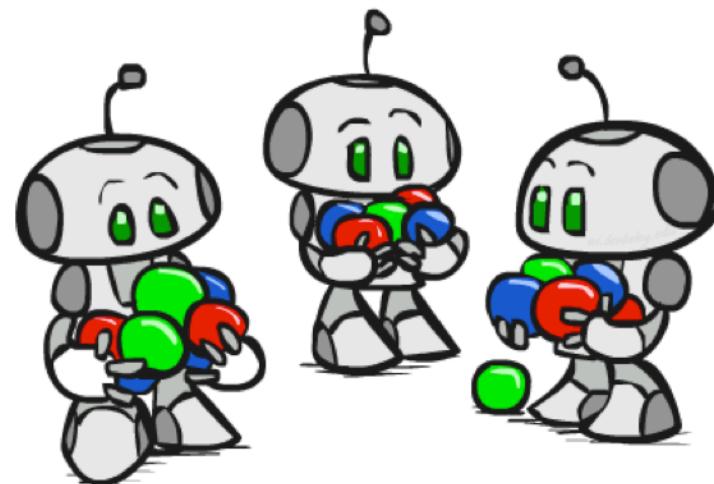


Initialization

- K-means is non-deterministic

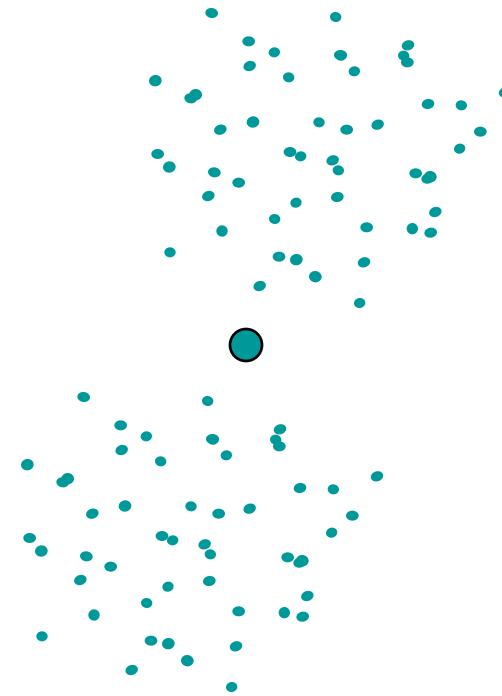
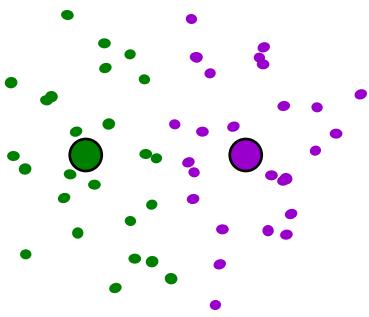
- Requires initial means
- It does matter what you pick!
- What can go wrong?

- Various schemes for preventing this kind of thing: variance-based split / merge, initialization heuristics

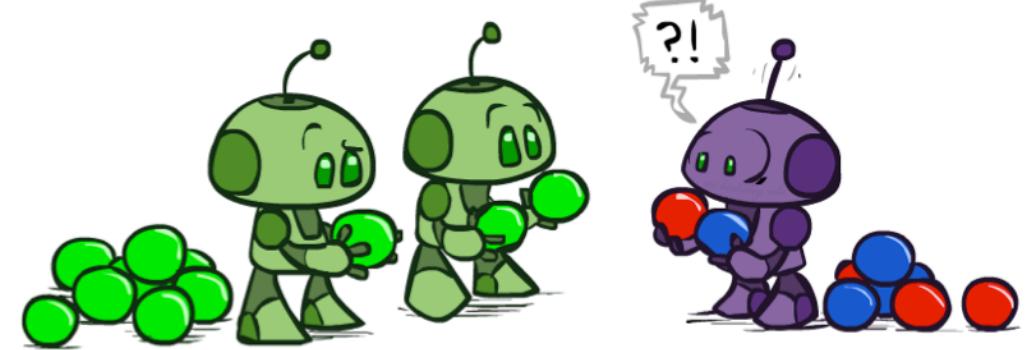


K-Means Getting Stuck

- A local optimum:



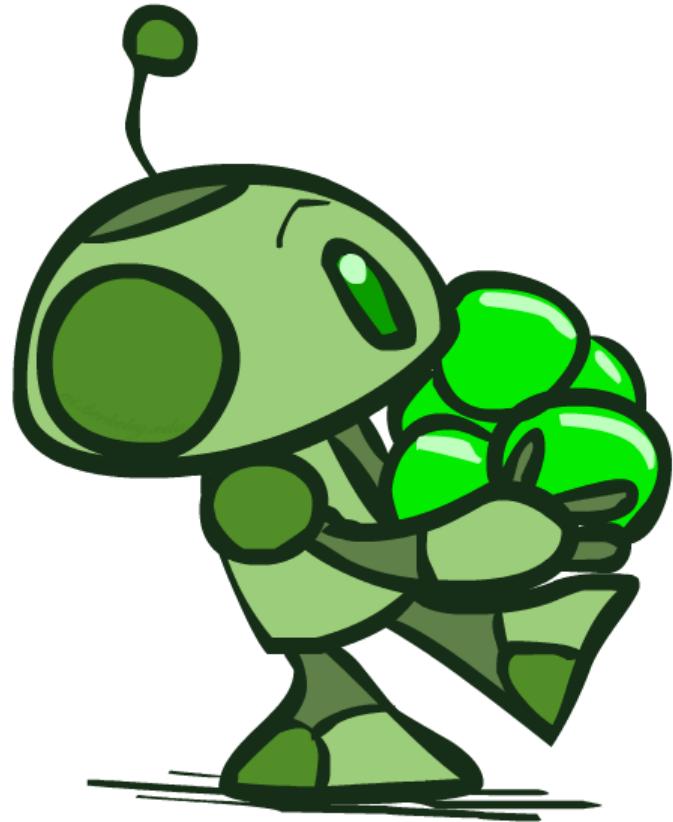
Why doesn't this work out like the earlier example, with the purple taking over half the blue?



K-Means Questions

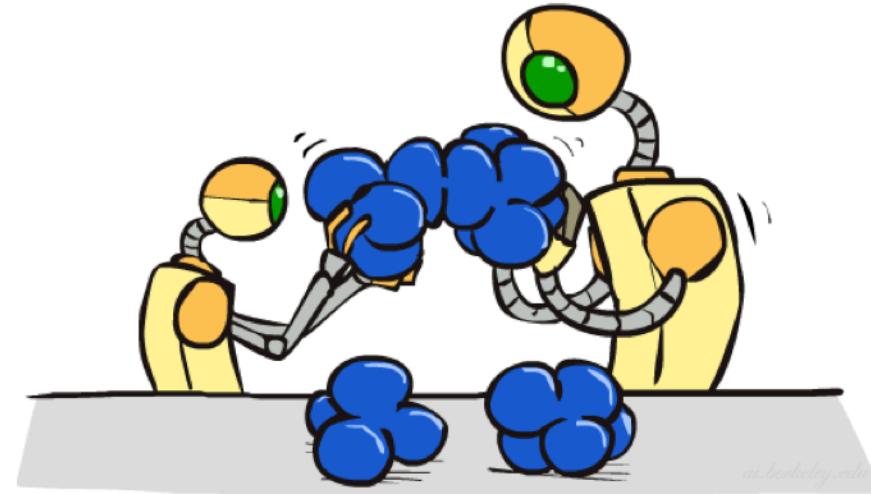
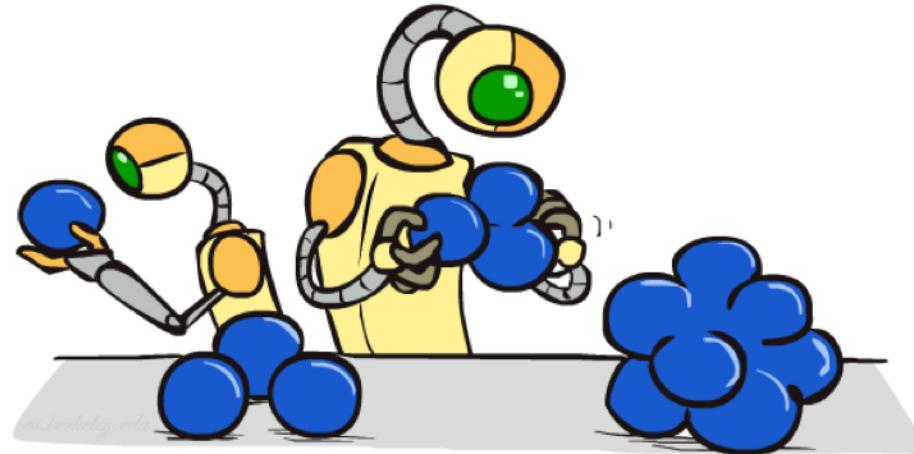
- Will K-means converge?
 - To a global optimum?
- Will it always find the true patterns in the data?
 - If the patterns are very very clear?
- Will it find something interesting?
- Do people ever use it?
- How many clusters to pick?

اگر الگو (pattern) خیلی واضح باشد
می تواند آن را پیدا کند .



Agglomerative Clustering

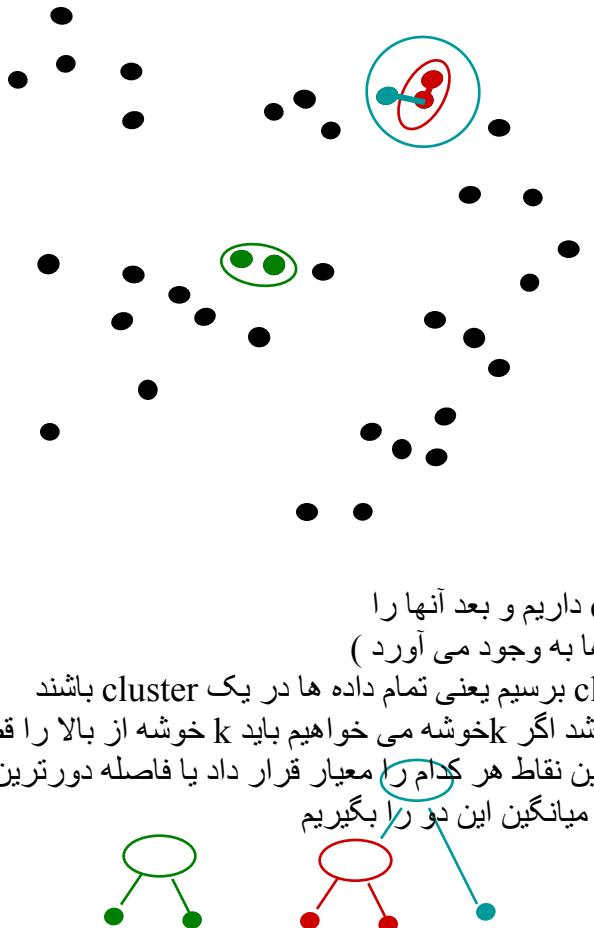
برای زمانی که اطلاعاتی از تعداد cluster ها نداریم



Agglomerative Clustering

- Agglomerative clustering:
 - First merge very similar instances
 - Incrementally build larger clusters out of smaller clusters
- Algorithm:
 - Maintain a set of clusters
 - Initially, each instance in its own cluster
 - Repeat:
 - Pick the two **closest** clusters
 - Merge them into a new cluster
 - Stop when there's only one cluster left
- Produces not one clustering, but a family of clusterings represented by a **dendrogram**

در این روش باید kn^2 تا فاصله را محاسبه کنیم بنابراین پیچیدگی محاسباتی بیشتری داریم



این روش بر عکس روش kmeans عمل می کند

kmeans : top down method
agglomerative : bottom up method

روش k means از روی نماینده ها خوشه تصمیم می گیرد
اما روش Agglomerative از روی داده شروع می کند به ساختن cluster ها
یعنی به خواص به نقاط داده خیلی توجه می کنه
و از روی داده خوشه هارا می سازد

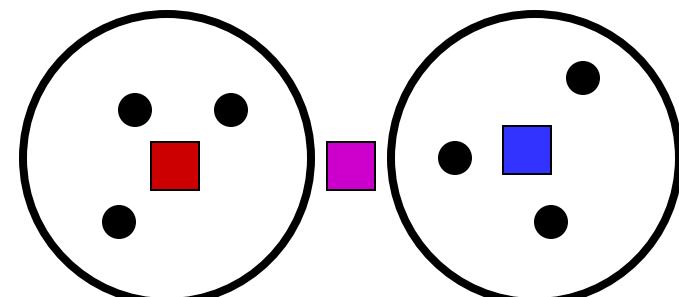
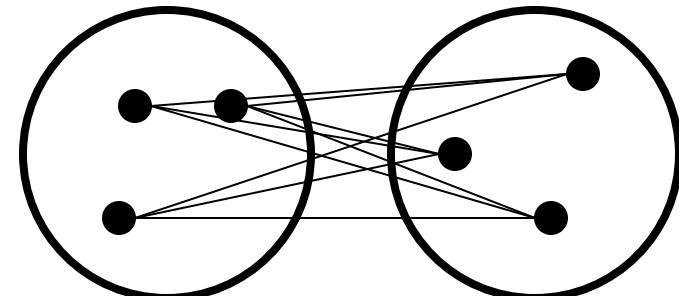
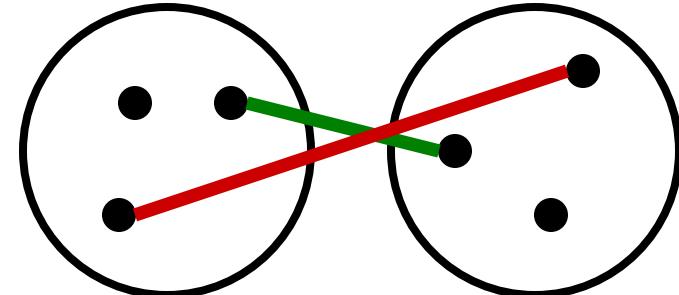
Agglomerative Clustering Algorithm :
تعداد زیادی نقطه داریم از بین همه این ها دو تا که
به هم نزدیک ترند را در یک cluster قرار می دهد

در واقع فرض می کنیم در ابتدا به اندازه تمام داده ها cluster داریم و بعد آنها را
با یک دیگر ادغام می کنیم (یک سلسله مراتب در cluster ها به وجود می آورد)

این کار را تا جایی ادامه می دهد که در نهایت به یک cluster برسیم یعنی تمام داده ها در یک cluster باشند
در نهایت بعد از این که سلسله مراتب برای خوشه ها ساخته شد اگر k خوشه می خواهیم باید k خوشه از بالا را قطع کنیم
برای مقایسه نزدیکی دو خوشه می توان یا فاصله نزدیک ترین نقاط هر کدام را میان فاصله دورترین نقاط یا
مجموع فواصل هر نقطه از هر خوشه یا بعد از تعیین نماینده میانگین این دو را بگیریم

Agglomerative Clustering

- How should we define “closest” for clusters with multiple elements?
- Many options
 - Closest pair (single-link clustering)
 - Farthest pair (complete-link clustering)
 - Average of all pairs
 - Ward’s method (min variance, like k-means)
- Different choices create different clustering behaviors



Example: Google News

Google™ Search News Search the Web Advanced news search Preferences

Search and browse 25,000 news sources updated continuously.

World » [edit](#) [X](#)

Heavy Fighting Continues As Pakistan Army Battles Taliban
Voice of America - 10 hours ago

By Barry Newhouse Pakistan's military said its forces have killed 55 to 60 Taliban militants in the last 24 hours in heavy fighting in Taliban-held areas of the northwest. [Pakistani troops battle Taliban militants for fourth day](#) guardian.co.uk

[Army: 55 militants killed in Pakistan fighting](#) The Associated Press
[Christian Science Monitor](#) - CNN International - Bloomberg - New York Times
[all 3,824 news articles »](#)

Sri Lanka admits bombing safe haven
guardian.co.uk - 3 hours ago

Sri Lanka has admitted bombing a "safe haven" created for up to 150000 civilians fleeing fighting between Tamil Tiger fighters and the army.
[Chinese billions in Sri Lanka fund battle against Tamil Tigers](#) Times Online
[Huge Humanitarian Operation Under Way in Sri Lanka](#) Voice of America
[BBC News](#) - Reuters - AFP - Xinhua
[all 2,492 news articles »](#)

Business » [edit](#) [X](#)

Buffett Calls Investment Candidates' 2008 Performance Subpar
Bloomberg - 2 hours ago

By Hugh Son, Erik Holm and Andrew Frye May 2 (Bloomberg) -- Billionaire Warren Buffett said all of the candidates to replace him as chief investment officer of Berkshire Hathaway Inc. failed to beat the 38 percent decline of the Standard & Poor's 500 ...
[Buffett offers bleak outlook for US newspapers](#) Reuters
[Buffett: Limit CEO pay through embarrassment](#) MarketWatch
[CNBC](#) - The Associated Press - guardian.co.uk
[all 1,454 news articles »](#) 

Chrysler's Fall May Help Administration Reshape GM
New York Times - 5 hours ago

Auto task force members, from left: Treasury's Ron Bloom and Gene Sperling, Labor's Edward Montgomery, and Steve Rattner. BY DAVID E. SANGER and BILL VLASIC WASHINGTON - Fresh from pushing Chrysler into bankruptcy, President Obama and his economic team ...

[Comment by Gary Chaison](#) Prof. of Industrial Relations, Clark University
[Bankruptcy reality sets in for Chrysler, workers](#) Detroit Free Press
[Washington Post](#) - Bloomberg - CNNMoney.com
[all 11,028 news articles »](#)   - 

U.S. » [edit](#) [X](#)

Weekend Opinionator: Souter, Specter and the Future of the GOP
New York Times - 48 minutes ago

By Tobin Harshaw An odd week. While Barack Obama celebrated his 100th day in office, the headlines were pretty much dominated by the opposition party, albeit not in the way many Republicans would have liked.
[US Supreme Court Vacancy An Early Test For Sen Specter](#) Wall Street Journal
[Letters: Arlen Specter, Notre Dame, Chrysler](#) Houston Chronicle
[The Associated Press](#) - Kansas City Star - Philadelphia Inquirer - Bangor Daily News
[all 401 news articles »](#)

Joe Biden, the Flu and You
New York Times - 48 minutes ago

By GAIL COLLINS The swine flu scare has made it clear why Barack Obama picked Joe Biden for vice president. David Brooks and Gail Collins talk between columns.
[After his flu warning, Biden takes the train home](#) The Associated Press
[Biden to visit Balkan states in mid-May](#) Washington Post
[AFP](#) - Christian Science Monitor - Bizjournals.com - Voice of America
[all 1,506 news articles »](#)

Top-level categories:
supervised classification

Story groupings:
unsupervised clustering