Linear Algebra Methods for Data Mining

Saara Hyvönen, Saara.Hyvonen@cs.helsinki.fi

Spring 2007

Lecture 3: QR, least squares, linear regression

QR decomposition

- Any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, $m \geq n$, can be transformed to upper triangular form by an orthogonal matrix:

$$\mathbf{A} = \mathbf{Q} \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix}$$

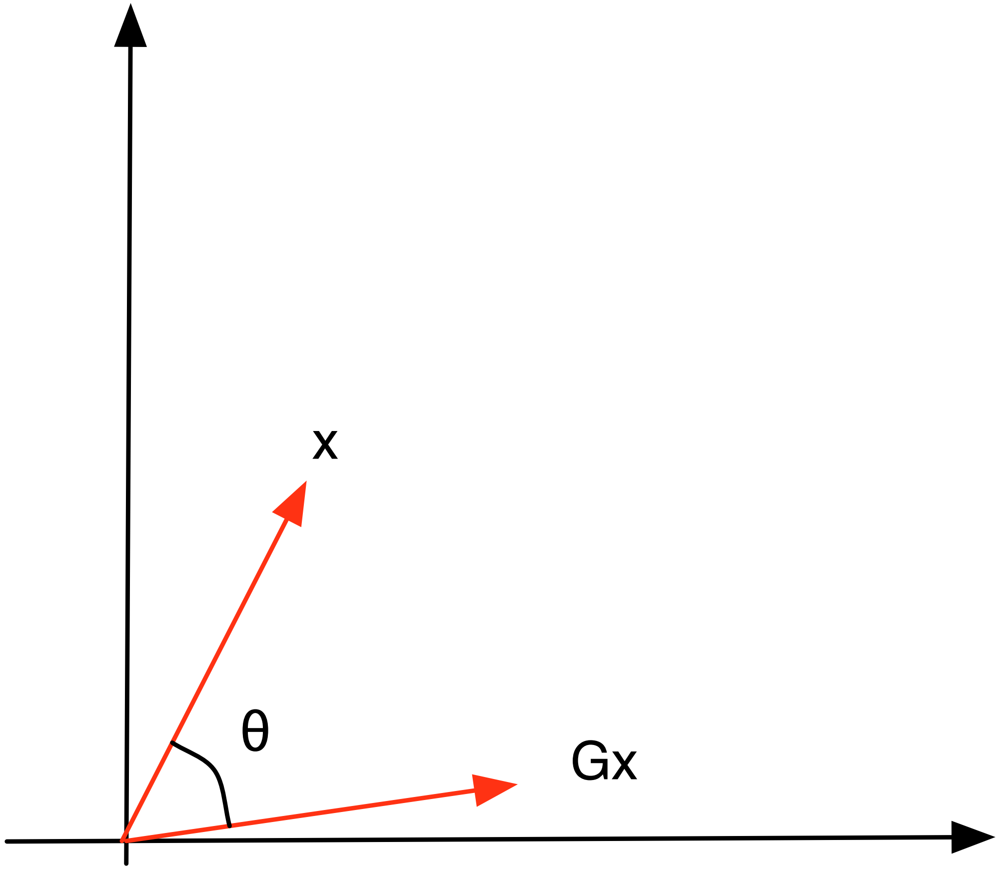
- If the columns of $\mathbf{A}$ are linearly independent, then $\mathbf{R}$ is non-singular.

How? By using the Givens rotation:

Let $\mathbf{x}$ be a vector. The parameters c and s, $c^2 + s^2 = 1$, can be chosen so that multiplication of $\mathbf{x}$ by

$$\mathbf{G} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & c & 0 & s \\ 0 & 0 & 1 & 0 \\ 0 & -s & 0 & c \end{pmatrix}$$

will zero the element 4 in vector $\mathbf{x}$ by a rotation in plane (2,4). How?


x
θ
Gx

"Skinny" QR decomposition

Partition $\mathbf{Q} = (\mathbf{Q}_1\ \mathbf{Q}_2)$, where $\mathbf{Q}_1 \in \mathbb{R}^{m \times n}$:

$$\mathbf{A} = (\mathbf{Q}_1\ \mathbf{Q}_2) \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{pmatrix} = \mathbf{Q}_1 \mathbf{R}_1.$$

```
A =  1     1     1
     1     2     4
     1     3     9
     1     4    16

[Q,R]=qr(A)  %the QR decomposition

Q =
   -0.5000    0.6708    0.5000    0.2236
   -0.5000    0.2236   -0.5000   -0.6708
   -0.5000   -0.2236   -0.5000    0.6708
   -0.5000   -0.6708    0.5000   -0.2236
R =
   -2.0000   -5.0000  -15.0000
         0   -2.2361  -11.1803
         0         0    2.0000
         0         0         0
```

```
[Q,R]=qr(A,0)    %the skinny version of QR

Q =
   -0.5000    0.6708    0.5000
   -0.5000    0.2236   -0.5000
   -0.5000   -0.2236   -0.5000
   -0.5000   -0.6708    0.5000


R =
   -2.0000   -5.0000  -15.0000
         0   -2.2361  -11.1803
         0         0    2.0000
```

Uniqueness?

Suppose $\mathbf{A} \in \mathbb{R}^{m \times n}$ has full column rank (i.e. the column vectors are linearly independent). The "skinny" QR factorization

$$\mathbf{A} = \mathbf{Q}_1 \mathbf{R}_1$$

is unique where $\mathbf{Q}_1$ has orthonormal columns and $\mathbf{R}_1$ is upper triangular with positive diagonal entries.

What is QR good for?

- It will come up when we discuss the computation of some matrix decompositions

- It can also be used for solving the least squares problem.

- Other applications exist.

Least squares for linear regression

- Consider problems of the following form: given a number of measurements $\mathbf{X} = [\mathbf{x}_1 ... \mathbf{x}_n]$ and an outcome $\mathbf{y}$, build a linear model

$$\hat{\mathbf{y}} = b_0 + \sum_{j=1}^{n} \mathbf{x}_j b_j,$$

  which uses the measurements $\mathbf{X}$ to predict the outcome $\mathbf{y}$.

- $\mathbf{X}$ = clinical measures of patients, $\mathbf{y}$ = level of cancer specific antigen.

- $\mathbf{X}$ = atmospheric measurements of each day, $\mathbf{y}$ = occurence of spontaneous particle formation.

Least squares for linear regression

- The model $\hat{\mathbf{y}} = b_0 + \sum_{j=1}^{n} \mathbf{x}_j b_j$ can be written in the form

$$\hat{\mathbf{y}} = \mathbf{X}^T \mathbf{b}, \quad \mathbf{b} = (b_1 \, ... \, b_n \, b_0)^T.$$

  To do this, one must append $\mathbf{X} = [\mathbf{x}_1 \, ... \, \mathbf{x}_n \; \mathbf{x}_{n+1}]$, where $\mathbf{x}_{n+1}$ is a vector of ones.

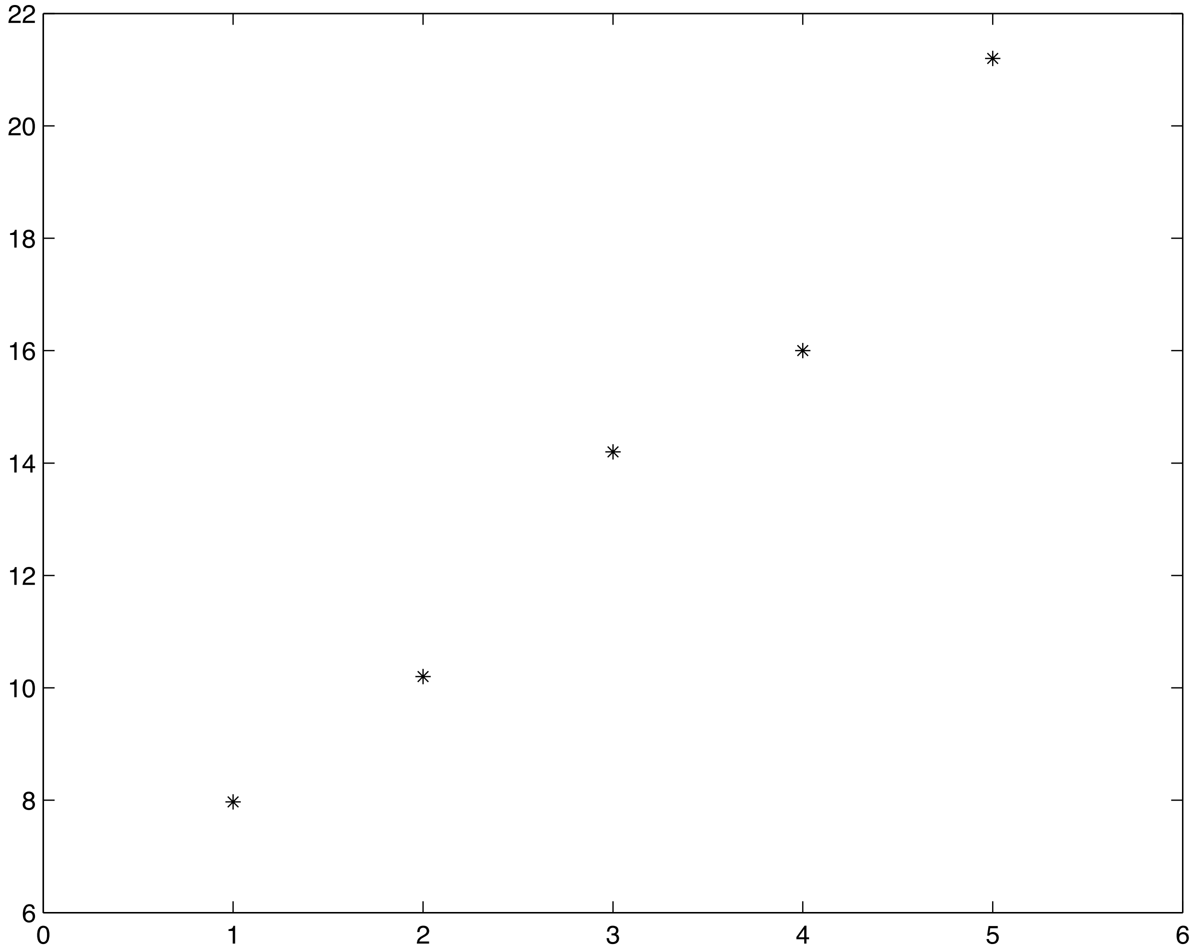
- Fitting a linear model to data is usually done using the method of least squares.

- Note: $\mathbf{X}$ need not be a square matrix!

Example

We have measurement data:

x	1	2	3	4	5
y	7.97	10.2	14.2	16.0	21.2


22
20
18
16
14
12
10
8
6
0
1
2
3
4
5
6

Example (continued)

x	1	2	3	4	5
y	7.97	10.2	14.2	16.0	21.2

We wish to find α and β such that $\alpha x + \beta = y$. Thus

$$
\begin{aligned}
\alpha + \beta &= 7.97 \\
2\alpha + \beta &= 10.2 \\
3\alpha + \beta &= 14.2 \\
4\alpha + \beta &= 16.0 \\
5\alpha + \beta &= 21.2
\end{aligned}
$$

Example (continued)

In matrix form:

$$\begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \\ 5 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 7.97 \\ 10.2 \\ 14.2 \\ 16.0 \\ 21.2 \end{pmatrix}$$

Overdetermined! (More equations than unknowns.)

Solve using the least squares method.

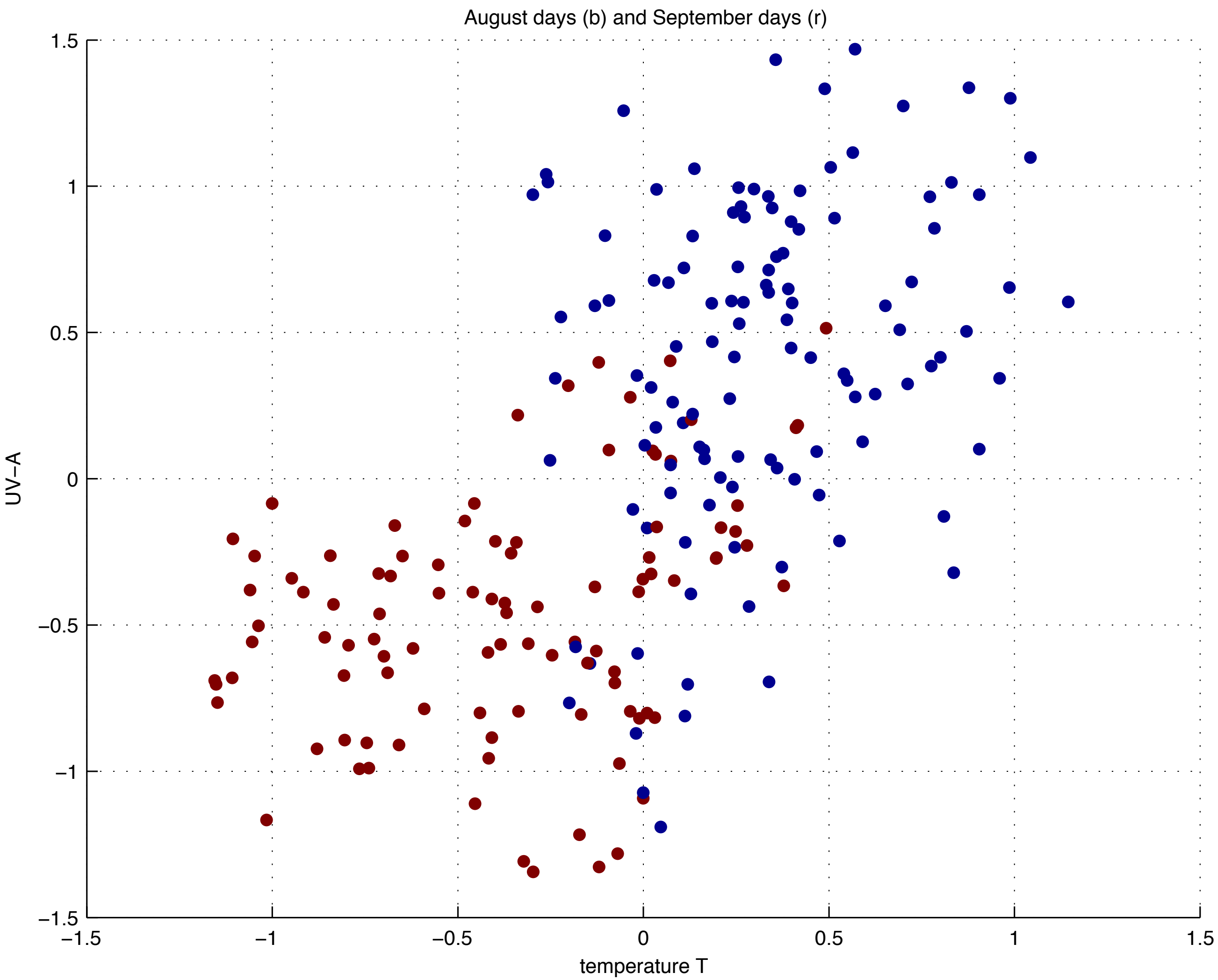
Example

- $\mathbf{X}$ = atmospheric measurements of each day: temperature, UV-A:

$$\mathbf{X} = \begin{pmatrix} t_1 & t_2 & \ldots & t_n \\ r_1 & r_2 & \ldots & r_n \end{pmatrix} \quad = \begin{pmatrix} \text{temperatures} \\ \text{UV-A measurements} \end{pmatrix}$$

- $\mathbf{y}$ = which month each day belongs to. Choices are: August ($y = 0$) or September ($y = 1$).

- Looking for $\mathbf{b}$ such that $\hat{\mathbf{y}} = \mathbf{X}^T\mathbf{b}$ and $\hat{\mathbf{y}} \approx \mathbf{y}$. Use method of least squares.


August days (b) and September days (r)
UV−A
temperature T
1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
1.5

The least squares problem

- Let $\mathbf{A} \in \mathbb{R}^{m \times n}$, $m \geq n$. The system

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

is called **overdetermined**: more equations than unknown. Usually such a system has no solution.

Example

$m = 3$, $n = 2$.

What to do?

- make the **residual vector**

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$$

  as small as possible. But how?

- Make $\mathbf{r}$ orthogonal to the columns of $\mathbf{A}$:

$$\mathbf{r}^T \begin{pmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \end{pmatrix} = \mathbf{r}^T \mathbf{A} = 0.$$

- Now write $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$ to get the **normal equations**

$$\mathbf{A}^T \mathbf{A}\mathbf{x} = \mathbf{A}^T \mathbf{b}; \quad \text{solve for } \mathbf{x}.$$

Example

```
A=      1   1                       b= 7.97
        2   1                          10.2
        3   1                          14.2
        4   1                          16.0
        5   1                          21.2

C=A'*A                  %Normal equations

C=     55     15
       15      5

x=C\(A'*b)

x=     3.2260
       4.2360
```

- If the column vectors of $\mathbf{A}$ are linearly independent, then the normal equations

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$$

are non-singular and have a unique solution

- BUT: the normal equations have two significant drawbacks:
  - forming $\mathbf{A}^T\mathbf{A}$ leads to loss of information.
  - the condition number of $\mathbf{A}^T\mathbf{A}$ is the square of that of $\mathbf{A}$:

$$\kappa(\mathbf{A}^T\mathbf{A}) = (\kappa(\mathbf{A}))^2$$

Example

```
 A =
     1     1
     2     1
     3     1
     4     1
     5     1

cond(A) =    8.3657

cond(A'*A) =   69.9857
```

Worse example

```
A =
      101    1
      102    1
      103    1
      104    1
      105    1

cond(A) =    7.5038e+03

cond(A'*A) =    5.6307e+07
```

Solving least squares problem using QR

$$\|\mathbf{r}\|^2 = \|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2 = \|\mathbf{b} - \mathbf{Q}\begin{pmatrix}\mathbf{R}\\ \mathbf{0}\end{pmatrix}\mathbf{x}\|^2 = \|\mathbf{Q}(\mathbf{Q}^T\mathbf{b} - \begin{pmatrix}\mathbf{R}\\ \mathbf{0}\end{pmatrix}\mathbf{x})\|^2$$

$$= \|\mathbf{Q}^T\mathbf{b} - \begin{pmatrix}\mathbf{R}\\ \mathbf{0}\end{pmatrix}\mathbf{x}\|^2$$

Partition $\mathbf{Q} = (\mathbf{Q}_1\ \mathbf{Q}_2)$, where $\mathbf{Q}_1 \in \mathbb{R}^{m\times n}$, and denote

$$\mathbf{Q}^T\mathbf{b} = \begin{pmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{pmatrix} := \begin{pmatrix}\mathbf{Q}_1^T\mathbf{b}\\ \mathbf{Q}_2^T\mathbf{b}\end{pmatrix}.$$

Then

$$\|\mathbf{r}\|^2 = \| \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} - \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix} \mathbf{x}) \|^2 = \|\mathbf{b}_1 - \mathbf{R}\mathbf{x}\|^2 + \|\mathbf{b}_2\|^2.$$

Minimize $\|\mathbf{r}\|$ by making the first term equal to zero: i.e. solve

$$\mathbf{R}\mathbf{x} = \mathbf{b}_1.$$

LS by QR

Theorem. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ have full column rank and a thin QR-decomposition $\mathbf{A} = \mathbf{Q}_1\mathbf{R}$. Then the least squares problem

$$\min_x \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$$

has the unique solution

$$\mathbf{x} = \mathbf{R}^{-1}\mathbf{Q}_1^T\mathbf{b}.$$

Example

```
A=      1   1                      b= 7.97
        2   1                         10.2
        3   1                         14.2
        4   1                         16.0
        5   1                         21.2
[Q1,R]=qr(A,0) %skinny QR

Q1 = -0.1348   -0.7628
   -0.2697   -0.4767
   -0.4045   -0.1907
   -0.5394    0.0953
   -0.6742    0.3814
```

```
R =   -7.4162   -2.0226
            0   -0.9535

x=R\(Q1'*b)

x =   3.2260
      4.2360
```

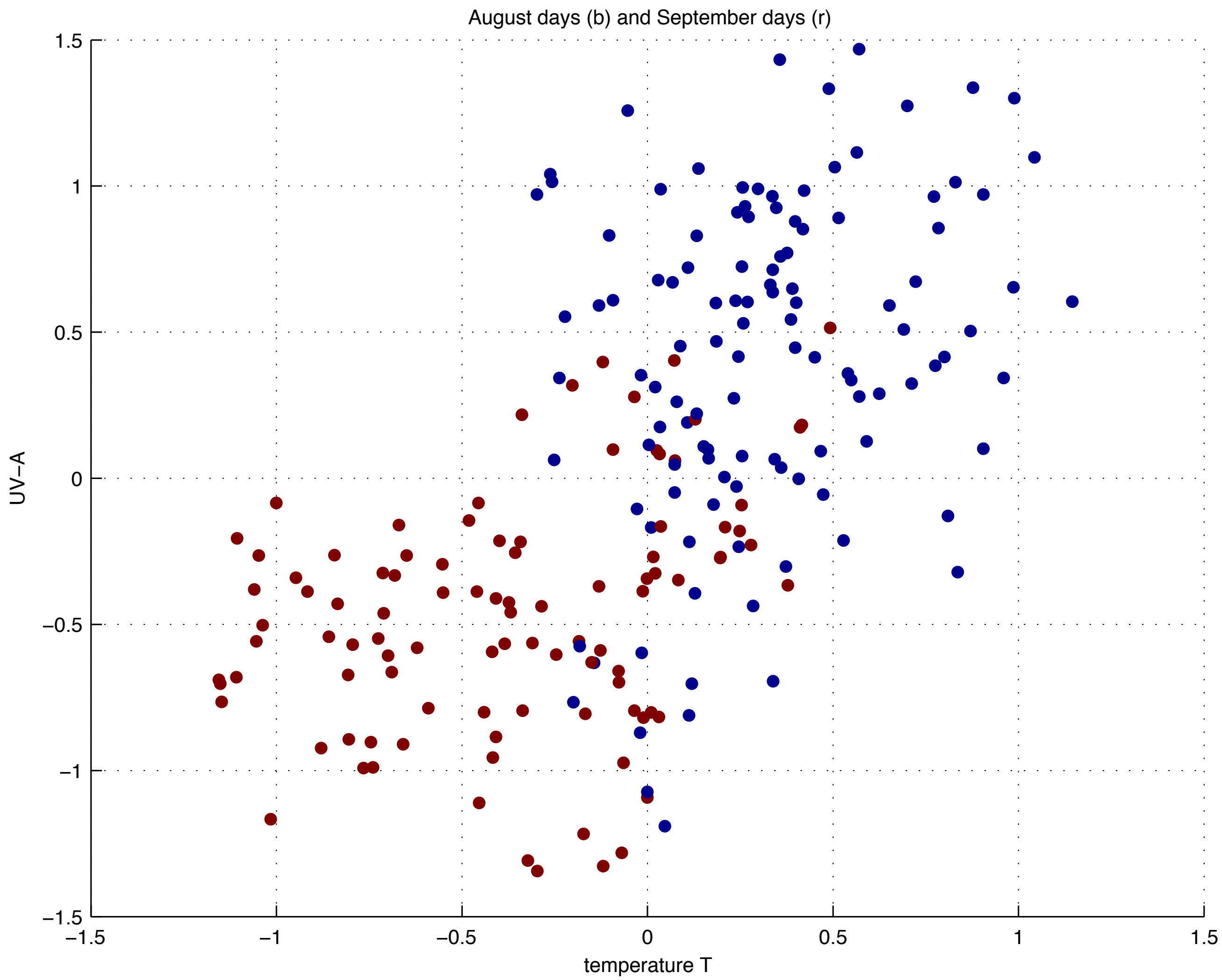
Back to this example:

- $\mathbf{X}$ = atmospheric measurements of each day: temperature, UV-A:

$$\mathbf{X} = \begin{pmatrix} t_1 & t_2 & \dots & t_n \\ r_1 & r_2 & \dots & r_n \end{pmatrix} \quad = \begin{pmatrix} \text{temperatures} \\ \text{UV-A measurements} \end{pmatrix}$$

- $\mathbf{y}$ = which month each day belongs to. Choices are: August ($y = 0$) or September ($y = 1$).

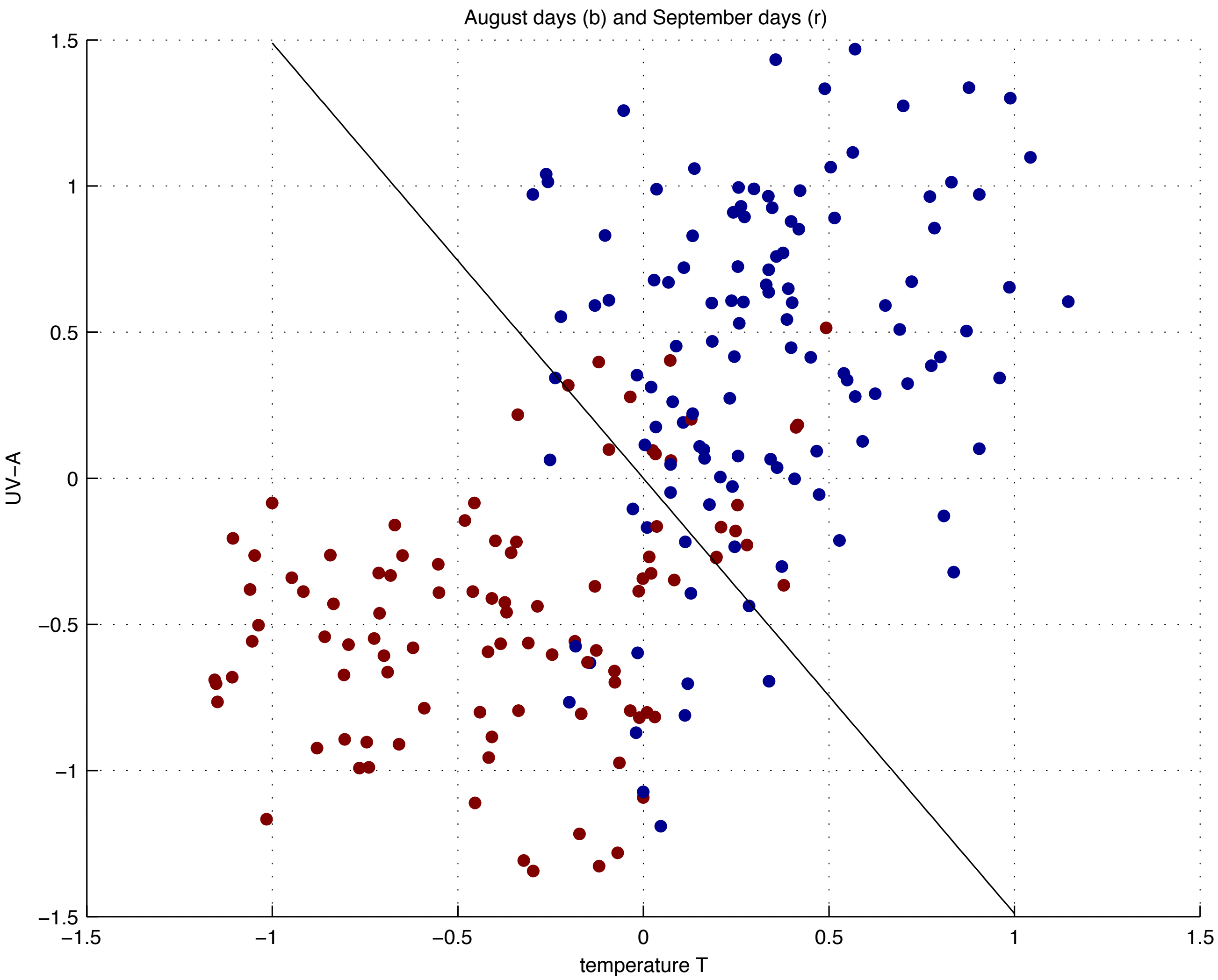
- Looking for $\mathbf{b}$ such that $\hat{\mathbf{y}} = \mathbf{X}^T\mathbf{b}$ and $\hat{\mathbf{y}} \approx \mathbf{y}$. Use method of least squares.


August days (b) and September days (r)
1.5
1
0.5
0
−0.5
−1
−1.5
UV−A
−1.5
−1
−0.5
0
0.5
1
1.5
temperature T

Example

```
 %X (transpose of) data matrix, 1st col: temperature, 2nd col: UV-A
 %y month vector, y(j)=0 if August and 1 if September

[length(find(y==0)) length(find(y==1))] =   116    99
                               %number of August and September days

size(X) =   215     2   %size of data matrix
Xap=[X ones(215,1)];

[Q1,R]=qr(Xap,0);

b=R\(Q1'*y) %this is the same as b=inv(R)*(Q1'*y)
b=   -0.4355
     -0.2923
      0.4605
```

```
yhat=Xap*b;  %least squares solution
[min(yhat) max(yhat)] =   -0.3506    1.2436

alpha=0.5;        %day classified as September if x'y>alpha
length(find(1.0*(yhat>alpha)-y)) =    34        %misclassifications
x=-1:0.01:1;db=(alpha-b(3)-b(1)*x)/b(2);%decision boundary x'b=alpha

%attn: in reality, choose alpha with more care!

scatter(Xnorm(:,1),Xnorm(:,2),20,y,'filled')
hold;plot(x,db,'k')
```


August days (b) and September days (r)
1.5
1
0.5
0
−0.5
−1
−1.5
UV−A
−1.5
−1
−0.5
0
0.5
1
1.5
temperature T

Updating the solution of the LS problem

Assume we have reduced the matrix and the right hand side

$$(\mathbf{A} \quad \mathbf{b}) \rightarrow \mathbf{Q}^T(\mathbf{A} \quad \mathbf{b}) = \begin{pmatrix} \mathbf{R} & \mathbf{Q}_1^T\mathbf{b} \\ \mathbf{0} & \mathbf{Q}_2^T\mathbf{b} \end{pmatrix}.$$

From this the solution of the LS problem is readily available.

Assume we have not saved $\mathbf{Q}$.

We then get a new observation $(\mathbf{a} \quad b)$, $\mathbf{a} \in \mathbb{R}^n$, $b \in \mathbb{R}$.

Do we have to recompute the whole solution?

Updating the solution of the LS problem

No. Instead, write the reduction on the previous slide as

$$\begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{a}^T & b \end{pmatrix} \rightarrow \begin{pmatrix} \mathbf{Q}^T & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{a}^T & b \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \mathbf{Q}_1^T\mathbf{b} \\ \mathbf{0} & \mathbf{Q}_2^T\mathbf{b} \\ \mathbf{a}^T & b \end{pmatrix}.$$

And reduce this to triangular form using plane rotations.

References

[1] Lars Eldén: Matrix Methods in Data Mining and Pattern Recognition, SIAM 2007.

[2] G. H. Golub and C. F. Van Loan. Matrix Computations. 3rd ed. Johns Hopkins Press, Baltimore, MD., 1996.

[3] T. Hastie, R. Tibshirani, J. Friedman: The Elements of Statistical Learning. Data mining, Inference and Prediction, Springer Verlag, New York, 2001.