Data Representation

AW

Lecture Overview

Moments

Discrete case:

Let a discrete attribute X be distributed with pmf $\Pr_X(x)$. The expectation of X denoted $E(X)$ is given by

$$E(X) = \sum_{a \in dom(X)} a \Pr(X = a)$$

Continuous case:

Let a continuous attribute X be distributed with pdf $f_X(x)$. The expectation of X denoted either $E(X)$ is given by

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

Sample mean For $s_X = \{x_1, \ldots, x_n\}$ sample mean μ_{S_X} is

$$\mu_{S_x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n}\left(\vec{1}_n \bullet \vec{s}_X\right)$$

Sample median m_{s_X} is a median of empirical cdf, i.e. if observed values s_X are sorted (i.e. sequence $x_{i_1}, \ldots, x_{i_n}$ of elements of s_x is such $x_{i_j} < x_{i_{j+1}}$ and whenever $i_j \neq i_k$ it holds $x_{i_j} \neq x_{i_k}$), then $m_{s_x} = x_{\lfloor \frac{n}{2} \rfloor + 1}$ if n odd.

More Moments

Variance $var(X)$ of a random variable X is the expectation of the square centered variable $X_c = X - E(X)$, i.e.

$$var(X) = E(X_c^2) = E\left((X - E(X))^2\right) = E(X^2 - 2XE(X) + E(X)^2) = E(X^2) - 2E(X)E(X) + E(X)^2 = E(X^2) - E(X)^2.$$

σ-sample variance is $\hat{\sigma}_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{s_X})^2 = \frac{\vec{s}_X^c \bullet \vec{s}_X^c}{n} = \frac{\left\|\vec{s}_X^c\right\|^2}{n}$ (biased)

s-sample variance is $\hat{s}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_{s_X})^2 = \frac{\vec{s}_X^c \bullet \vec{s}_X^c}{n-1} = \frac{\left\|\vec{s}_X^c\right\|^2}{n-1}$ (unbiased)

Standard deviation $\sigma(X)$ (or σ_X) of random variable X is square root of its variance.

Lecture Overview

Covariance

Covariance of two random variables X and Y is the expectation of a product of these variables centered, i.e.

$$cov(X, Y) = E\left(X^c \cdot Y^c\right) = E((X - E(X) \cdot (Y - E(Y))$$

Expanding we get

$$\begin{aligned} cov(X, Y) &= E((X - E(X) \cdot (Y - E(Y)) \\ &= E(XY - X \cdot E(Y) - Y \cdot E(X) + E(X) \cdot E(Y)) \\ &= E(XY) - E(X) \cdot E(Y) \end{aligned}$$

- If X and Y are independent then $E(XY) = E(X)E(Y)$ and $cov(X, Y) = 0$.
- Yet if $cov(X, Y) = 0$ we cannot claim that X and Y are independent.
  - All we can say is that there is no linear dependence between them, because if $Y = aX + b$ then $E(XY) \neq E(X) \cdot E(Y)$ since $\sigma_X^2 \neq 0$ for a random variable
  - But we cannot rule out that there might be a higher order relationship or dependence between the variables.

Covariance

Covariance of two random variables X and Y is the expectation of a product of these variables centered, i.e.

$$cov(X,Y) = E\left(X^c \cdot Y^c\right) = E((X - E(X) \cdot (Y - E(Y)) = E(XY) - E(X) \cdot E(Y)$$

Sample covariance. Given simultaneous samples $s_X = \{x_1, \ldots, x_n\}$ and $s_Y = \{y_1, \ldots, y_n\}$. Then sample covariance is

$cov(s_X, s_Y) = \frac{1}{n-1} \sum_{j=1}^{n} \sum_{i=1}^{n} (x_j - \mu_{s_X})(y_i - \mu_{s_Y}) = \frac{\vec{s}_X^c \bullet \vec{s}_Y^c}{n-1}$ (unbiased)

or

$cov(s_X, s_Y) = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{k} (x_j - \mu_{s_X})(y_i - (\mu_{s_Y}) == \frac{\vec{s}_X^c \bullet \vec{s}_Y^c}{n}$ (biased)

The latter should be used when population means are known.

Here $\vec{s}_X^c$ and $\vec{s}_Y^c$ are centered sample vectors.

Meaning of Covariance

- Covariance is a measure of how much two random variables change together
- If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the smaller values, the covariance is a positive number.
- If the variables tend to show opposite behavior (i.e. the greater values of one variable mainly correspond to the smaller values of the other and vice versa) the covariance is negative.
- The sign of the covariance shows the tendency in the linear relationship between the variables.

Lecture Overview

Example of Covariance Computation

For discrete random vaiables X, Y that have joint probability mass function $p(X, Y)$ we compute $cov(X, Y) = \sum_j \sum_i (x_j - E(X))(y_i - E(Y))p(X = x_j, Y = y_i)$

Example. Let the pmf be given by:

$$
\begin{array}{c} \\ X \end{array}
\begin{array}{c} Y \\ \left(\begin{array}{c|ccc|c} p(x,y) & 1 & 2 & 3 & p(x) \\ \hline 1 & 0.25 & 0.25 & 0 & 0.5 \\ 2 & 0 & 0.25 & 0.25 & 0.5 \\ \hline p(y) & 0.25 & 0.5 & 0.25 & 1 \end{array}\right) \end{array}
$$

Then $E(X) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} = \frac{3}{2}$ and $E(Y) = (1 + 3) \cdot \frac{1}{4} + 2 \cdot \frac{1}{2} = 2$. So,

$$
\begin{array}{c} \\ X^c \end{array}
\begin{array}{c} Y^c \\ \left(\begin{array}{c|ccc|c} p(x,y) & -1 & 0 & 1 & p(x) \\ \hline -0.5 & 0.25 & 0.25 & 0 & 0.5 \\ 0.5 & 0 & 0.25 & 0.25 & 0.5 \\ \hline p(y) & 0.25 & 0.5 & 0.25 & 1 \end{array}\right) \end{array}
$$

and $cov(X, Y) = (-.5) \cdot (-1 \cdot .25 + 0 \cdot .25 + 1 \cdot 0) + (.5)(-1 \cdot 0 + 0 \cdot .25 + 1 \cdot .25) = .25$

Correlation

Z-score of indiviual (raw) value of random variable X is the distance of this value from the mean measured in the number of standard deviations: $z_X = \frac{X - E(X)}{\sigma_X}$

Pearson correlation coefficient for random variables X and Y denoted by ρ_{XY} is the expectation of a product of Z-scores of X and Y, i.e.

$$\rho_{XY} = E\left(\frac{X - E(X)}{\sigma_X} \cdot \frac{Y - E(Y)}{\sigma_Y}\right) = \frac{cov(XY)}{\sigma_X \cdot \sigma_Y}$$

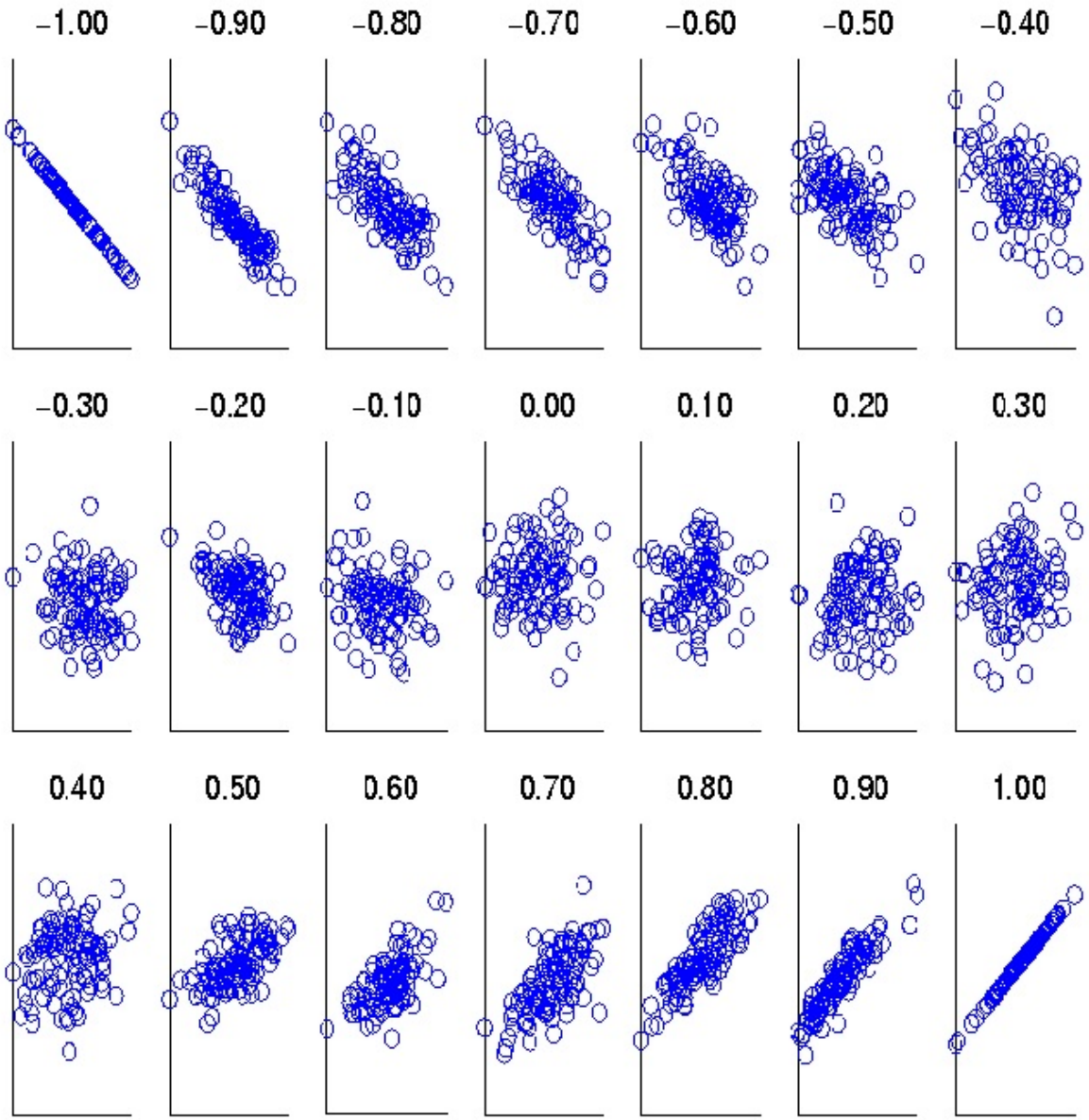
Sample correlation is given using sample variance and sample standard deviations. For sample vectors $\vec{s}_X,\ \vec{s}_Y \in \mathbb{R}^n$ drawn from random variables X and Y we have

$$\hat{\rho}_{XY} = \frac{\widehat{cov(XY)}}{\hat{\sigma}_X \cdot \hat{\sigma}_Y} = \frac{\frac{1}{n-1}\vec{s}_X^c \bullet \vec{s}_Y^c}{\sqrt{\frac{1}{n-1}\left\|\vec{s}_X^c\right\|^2} \cdot \sqrt{\frac{1}{n-1}\left\|\vec{s}_Y^c\right\|^2}} = \frac{\vec{s}_X^c \bullet \vec{s}_Y^c}{\left\|\vec{s}_X^c\right\| \cdot \left\|\vec{s}_Y^c\right\|}$$

Observe that $-1 \leq \hat{\rho}_{XY} \leq 1$ since by Cauchy-Schwartz inequality we have that $|\vec{s}_X^c \bullet \vec{s}_Y^c| \leq \|\vec{s}_X^c\| \cdot \|\vec{s}_Y^c\|$, so $\left|\frac{\widehat{cov(XY)}}{\hat{\sigma}_X \cdot \hat{\sigma}_Y}\right| = \frac{\left|\vec{s}_X^c \bullet \vec{s}_Y^c\right|}{\left\|\vec{s}_X^c\right\| \cdot \left\|\vec{s}_Y^c\right\|} \leq 1$

If X and Y are independent then $cov(X, Y) = 0$ and $\rho_{X,Y} = 0$. However, if $\rho_{X,Y} = 0$, we cannot claim that X and Y are independent.

Visually Evaluating Correlation



Meaning of Correlation

- Pearson correlation is +1 in the case of a perfect positive (increasing) linear relationship between X and Y.
- Pearson correlation is -1 in the case of a perfect decreasing (negative) linear relationship (anti-correlation)
- If the variables are independent, Pearson's correlation coefficient is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables.
- Value between -1 and 1 indicate a the degree oflinear (in)dependencebetween the variables
- Observe that $\hat{\rho}_{XY} = \frac{\vec{s}_X^c \bullet \vec{s}_Y^c}{\|\vec{s}_X^c\| \cdot \|\vec{s}_Y^c\|} = \frac{\vec{s}_X^c}{\|\vec{s}_X^c\|} \bullet \frac{\vec{s}_Y^c}{\|\vec{s}_Y^c\|} = \cos(\Theta_{\vec{s}_X^c \vec{s}_Y^c})$ so the correlation coefficient is simply the cosine of the angle between the two centered sample vectors.

Operating with (co)Variances

Let $X,\ Y,\ Z,\ W$ be random variables and $a, b \in \mathbb{R}$ are constants. Then

Rule 1 $\sigma^2_{a+bX} = b^2\sigma^2_X$

Rule 2 Variance of linear combination:

$\sigma^2_{aX+bY} = a^2\sigma^2_X + b^2\sigma^2_Y + ab\rho_{XY}\sigma^2_X\sigma^2_Y$

$\sigma^2_{aX-bY} = a^2\sigma^2_X + b^2\sigma^2_Y - ab\rho_{XY}\sigma^2_X\sigma^2_Y$

so when $\rho_{XY} = 0$, i.e. X, Y are uncorrelated (e.g. when they are indendent), we have $\sigma^2_{aX\pm bY} = a^2\sigma^2_X + b^2\sigma^2_Y$

Rule 3 Bilinearity of Covariance.

$cov(aX + bY, cW + dZ) =$

$ac \cdot cov(X, W) + ad \cdot cov(X, Z) + bc \cdot cov(Y, W) + bd \cdot cov(Y, Z)$

Lecture Overview

Data Normalizzation

To normalize the data is to make it consistent in some way. Two basic types of normalization:

- feature normalization is adjusting each value of a feature the same way across all examples
  - Typical feature normalizations are centering (i.e. $x_i \rightarrow x_i - \mu_X$), variance scaling (i.e. $x_i \rightarrow \frac{x_i}{\hat{\sigma}_X}$) and asolute scaling (i.e. $x_i \rightarrow \frac{x_i}{\max_i x_i}$)
- example normalization treats each example as a vector in some space (usually $\mathbb{R}^n$). Normalization is then linear automorfism of this space
- The main advantage to example normalization is that it makes comparisons more straightforward across data sets
  - Most common example normalization is vector normalization (i.e. $\vec{x}_i \rightarrow \frac{\vec{x}_i}{\|\vec{x}_i\|}$)

Irrelevant and Redundunt Features

- Intuitively, an irrelevant feature is one that is completely uncorrelated with the prediction. We'd like to prune irrelevant features.
  - It is easy in the case of binary features: if a binary feature only appears small number times in the data, you simply remove it from consideration.
  - How to extend the idea of does not occur much to real values?
  - A reasonable definition is to look for features with low variance.
- Intuitively, a redundant feature is one that is a function of another feature(s)
  - How to define redundunt features?
  - A reasoable definition is to look for features that are highly correlated with other features

Irrelevant and Redundunt Features

- Intuitively, an irrelevant feature is one that is completely uncorrelated with the prediction. We'd like to prune irrelevant features.
  - It is easy in the case of binary features: if a binary feature only appears small number times in the data, you simply remove it from consideration.
  - How to extend the idea of does not occur much to real values?
  - A reasonable definition is to look for features with low variance.
- Intuitively, a redundant feature is one that is a function of another feature(s)
  - How to define redundunt features?
  - A reasoable definition is to look for features that are highly correlated with other features

Irrelevant and Redundunt Features

- Intuitively, an irrelevant feature is one that is completely uncorrelated with the prediction. We'd like to prune irrelevant features.
  - It is easy in the case of binary features: if a binary feature only appears small number times in the data, you simply remove it from consideration.
  - How to extend the idea of does not occur much to real values?
  - A reasonable definition is to look for features with low variance.
- Intuitively, a redundant feature is one that is a function of another feature(s)
  - How to define redundunt features?
  - A reasoable definition is to look for features that are highly correlated with other features

Irrelevant and Redundunt Features

- Intuitively, an irrelevant feature is one that is completely uncorrelated with the prediction. We'd like to prune irrelevant features.
  - It is easy in the case of binary features: if a binary feature only appears small number times in the data, you simply remove it from consideration.
  - How to extend the idea of does not occur much to real values?
  - A reasonable definition is to look for features with low variance.
- Intuitively, a redundant feature is one that is a function of another feature(s)
  - How to define redundunt features?
  - A reasoable definition is to look for features that are highly correlated with other features

Tool - Sample Covariance Matrix

- Let data matrix D consists of m instances (rows) that each is a record of values of n features $X_1, \ldots, X_n$ (columns). We treat each instance as a sample of features.

  Example: 10×2 matrix D has 2 features - columns $\vec{X}_1, \vec{X}_2$. These features take values $x_{1,1}, \ldots x_{10,1}$ and $x_{1,2}, \ldots x_{10,2}$ respectively. An instance $\#3$ is then a pair $\langle x_{3,1}, x_{3,2} \rangle$. So the value of feature $\vec{X}_1$ in the instance $\#3$ is $x_{3,1}$ while the value of feature $\vec{X}_2$ in the instance $\#3$ is $x_{3,2}$.

- Let D^c be centered D matrix, i.e. $D^c = (\vec{X}_1^c, \ldots, \vec{X}_n^c)$ where $\vec{X}_i^c = \vec{X}_i - \left(\frac{\vec{X}_i \bullet \vec{1}}{n} \right) \cdot \vec{1}$.

  The matrix $\Sigma_D = \frac{1}{n-1} (D^c)^T D^c$ is called sample covariance matrix of D because its elements are sample variances and covariances of features. For example, element (i,j) is $\frac{1}{n-1} \left(\vec{X}_i^c \right)^T \vec{X}_j^c = \widehat{cov}_{X_i X_j}$ while element (i,i) is $\frac{1}{n-1} \left(\vec{X}_i^c \right)^T \vec{X}_i^c = \hat{\sigma}^2_{X_i}$. We write $\widehat{cov}_{ij}$ instead of $\widehat{cov}_{X_i X_j}$ and $\hat{\sigma}_i^2$ instead of $\hat{\sigma}^2_{X_i}$ in this conntext.

- Trace $tr(\Sigma_D)$ of the sample covariance matrix (i.e. sum of all diagonal elements) is total variance in data D (sum of variances of all attributes).

Lecture Overview

Covariance and Correlation in R

- Sample covariance of 2 features (sepal length and sepal width) in irises data

```
cov(iris$Sepal.Length,iris$Sepal.Width)
```

- Sample covariance matrix of all features in irises data

```
cov(iris[1:4])
```

- Sample correlation coefficient of 2 features (sepal length and sepal width) in irises data

```
cor(iris$Sepal.Length,iris$Sepal.Width)
```

- Correlation matrix of all features in irises data

```
cor(iris[1:4])
```

- Correlation and its visual ppresentation: function `pairs` plot columns of matrix against each other

```
pairs(iris[1:4],main = "Iris Data",pch = 21, bg =
 c("red","green","blue")[unclass(iris$Species)])
```

  - Parameter `pch` defines the form of a point -21 means circles, 23-squares, etc.
  - Parameter `bg` means background color of the plotting points (when there is background as in circles or squares, i.e. when 21=pch=25). It defines background color assigned to points depending on values `iris$Species` attribute to each of the 3 in order defined by frame are assigned colors r,g,b

Histograms and Quantiles in R

```
data(iris)
    #load dataset into memory
hist(iris[,2],probability=TRUE,main="Y")
    #historgram with frequencies (not raw counts
    #for raw counts change to probability =FALSE
dev.copy2pdf(file="histogram.pdf")
    #save plot
quantile(example[,1],c(seq(from=0,to=1,by=0.1)))
#compute quantiles with a step 0.1
plot(quantile(iris[,1],c(seq(from=0,to=1,by=0.02))),
    quantile(iris[,2],c(seq(from=0,to=1,by=0.02))),
    main="Sepal Length vs Sepal Width", xlab="sepal
    length by 0.1", ylab="sepal width by 0.02")
    #plot of quatile vs. quantile
    plot(ecdf(iris$Petal.Length))
    #sample cdf
```