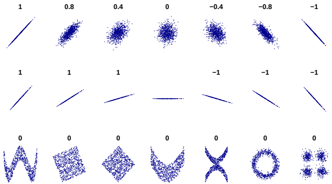
Interpretable Machine Learning

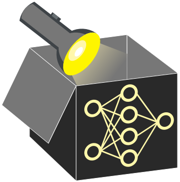
Intro to IML
Correlation and Dependencies



Learning goals

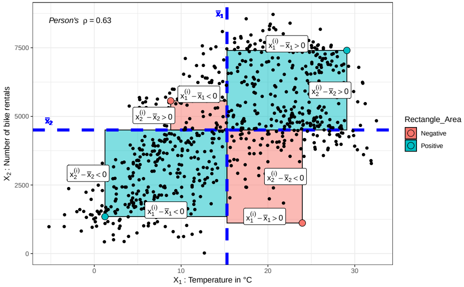
- Pearson correlation
- Coefficient of determination R^2
- Mutual information
- Correlation vs. dependence

PEARSON'S CORRELATION COEFFICIENT ρ



Correlation often refers to Pearson's correlation (measures only **linear relationship**)

$$\rho(X_1, X_2) = \frac{\sum_{i=1}^{n}(x_1^{(i)} - \bar{x}_1)\cdot(x_2^{(i)} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_1^{(i)} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_2^{(i)} - \bar{x}_2)^2}} \in [-1, 1]$$



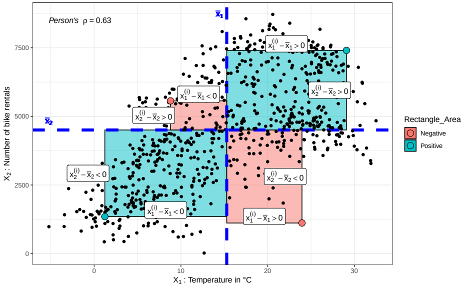
Geometric interpretation of ρ:

- Numerator is sum of rectangle's area with width $x_1^{(i)} - \bar{x}_1$ and height $x_2^{(i)} - \bar{x}_2$
- Areas enter numerator with positive (**+**) or negative (**-**) sign, depending on position
- Denominator scales the sum into the range $[-1, 1]$

PEARSON'S CORRELATION COEFFICIENT ρ

Correlation often refers to Pearson's correlation (measures only **linear relationship**)

$$\rho(X_1, X_2) = \frac{\sum_{i=1}^{n} (x_1^{(i)} - \bar{x}_1) \cdot (x_2^{(i)} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n} (x_1^{(i)} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_2^{(i)} - \bar{x}_2)^2}} \in [-1, 1]$$

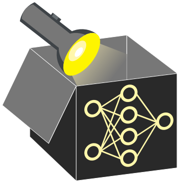


Geometric interpretation of ρ:

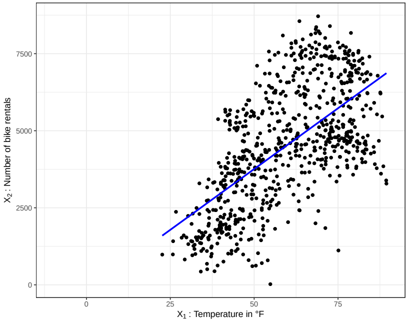
- Numerator is sum of rectangle's area with width $x_1^{(i)} - \bar{x}_1$ and height $x_2^{(i)} - \bar{x}_2$
- Areas enter numerator with positive (**+**) or negative (**-**) sign, depending on position
- Denominator scales the sum into the range $[-1, 1]$

- $\rho > 0$ if positive areas dominate negative areas
  $\rightsquigarrow X_1, X_2$ positive correlated
- $\rho < 0$ if negative areas dominate positive areas
  $\rightsquigarrow X_1, X_2$ negative correlated
- $\rho = 0$ if area of rectangles cancels out $\rightsquigarrow X_1, X_2$ linearly uncorrelated

COEFFICIENT OF DETERMINATION R^2



Another method to evaluate **linear dependency** between features is R^2



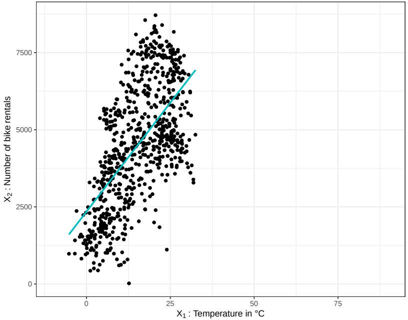
- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$

$\rightsquigarrow$ Slope $\theta_1 = 0 \Rightarrow$ no dependence

$\rightsquigarrow$ Large slope $\Rightarrow$ strong dependence

COEFFICIENT OF DETERMINATION R^2

Another method to evaluate **linear dependency** between features is R^2



- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$

$\leadsto$ Slope $\theta_1 = 0 \Rightarrow$ no dependence

$\leadsto$ Large slope $\Rightarrow$ strong dependence

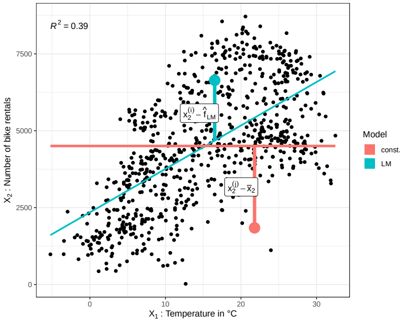
- Exact θ_1 score problematic

$\leadsto$ Re-scaling of x_1 or x_2 changes θ_1

$\leadsto$ °F $\longrightarrow$ °C $\Rightarrow \theta_1 = 78 \longrightarrow \theta_1^* = 141$

COEFFICIENT OF DETERMINATION R^2

Another method to evaluate **linear dependency** between features is R^2



- Fit a linear model:
  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$
  - ⇝ Slope $\theta_1 = 0 \Rightarrow$ no dependence
  - ⇝ Large slope $\Rightarrow$ strong dependence
- Exact θ_1 score problematic
  - ⇝ Re-scaling of x_1 or x_2 changes θ_1
- Set SSE_{LM} in relation to SSE of a constant model $\hat{f}_c = \bar{x}_2$

  $SSE_{LM} = \sum_{i=1}^{n}(x_2^{(i)} - \hat{f}_{LM}(x_1^{(i)}))^2$

  $SSE_c = \sum_{i=1}^{n}(x_2^{(i)} - \bar{x}_2)^2$

$\Rightarrow$ Measure of fitting quality of LM: $R^2 = 1 - \frac{SSE_{LM}}{SSE_c} \in [0, 1]$

$\Rightarrow \rho(X_1, X_2) = R$

JOINT, MARGINAL AND CONDITIONAL DISTRIBUTION

For two discrete random variables X_1, X_2:

Joint distribution

$$p_{X_1,X_2}(x_1, x_2) = \mathbb{P}(X_1 = x_1, X_2 = x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

JOINT, MARGINAL AND CONDITIONAL DISTRIBUTION

For two discrete random variables X_1, X_2:

Joint distribution

$$p_{X_1,X_2}(x_1, x_2) = \mathbb{P}(X_1 = x_1, X_2 = x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

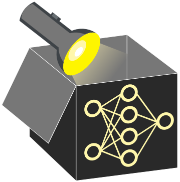
Marginal distribution

$$p_{X_1}(x_1) = \mathbb{P}(X_1 = x_1) = \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

$\rightsquigarrow$ In continuous case with integrals

JOINT, MARGINAL AND CONDITIONAL DISTRIBUTION



For two discrete random variables X_1, X_2:

Joint distribution

$$p_{X_1,X_2}(x_1, x_2) = \mathbb{P}(X_1 = x_1, X_2 = x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

Marginal distribution

$$p_{X_1}(x_1) = \mathbb{P}(X_1 = x_1) = \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2)$$

p_{X_1,X_2}	$\mathbb{P}(X_2 = 0)$	$\mathbb{P}(X_2 = 1)$	p_{X_1}
$\mathbb{P}(X_1 = 0)$	0.2	0.3	0.5
$\mathbb{P}(X_1 = 1)$	0.1	0.4	0.5
p_{X_2}	0.3	0.7	1

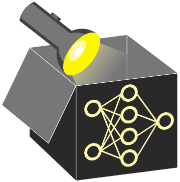
$\rightsquigarrow$ In continuous case with integrals

Conditional distribution

$$p_{X_1|X_2}(x_1|x_2) = \mathbb{P}(X_1 = x_1|X_2 = x_2)$$
$$= \frac{p_{X_1,X_2}(x_1, x_2)}{p_{X_2}(x_2)}$$

	$x_2 = 0$	$x_2 = 1$
$\mathbb{P}(X_1 = 0\|X_2 = x_2)$	0.67	0.43
$\mathbb{P}(X_1 = 1\|X_2 = x_2)$	0.33	0.57
$\sum$	1	1

DEPENDENCE

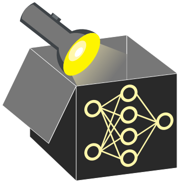


Dependence: Describes general dependence structure (e.g., non-lin. relationships)

- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

DEPENDENCE



Dependence: Describes general dependence structure (e.g., non-lin. relationships)

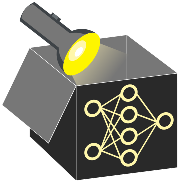
- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

- Equivalent definition (knowing X_k gives no info about X_j and vice versa):

  $\mathbb{P}(X_j | X_k) = \mathbb{P}(X_j)$ and $\mathbb{P}(X_k | X_j) = \mathbb{P}(X_k)$ (follows from cond. probability)

DEPENDENCE



Dependence: Describes general dependence structure (e.g., non-lin. relationships)

- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

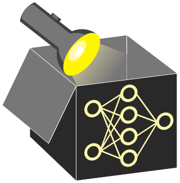
$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

- Equivalent definition (knowing X_k gives no info about X_j and vice versa):

  $\mathbb{P}(X_j|X_k) = \mathbb{P}(X_j)$ and $\mathbb{P}(X_k|X_j) = \mathbb{P}(X_k)$ (follows from cond. probability)

- Measuring complex dependencies is difficult but different measures exist Examples
  $\leadsto$ Spearman correlation (measures monotonic dependencies via ranks)
  $\leadsto$ Information-theoretical measures like mutual information
  $\leadsto$ Kernel-based measures like Hilbert-Schmidt Independence Criterion (HSIC)

DEPENDENCE



Dependence: Describes general dependence structure (e.g., non-lin. relationships)

- Definition: X_j, X_k independent $\Leftrightarrow$ joint distribution is product of marginals:

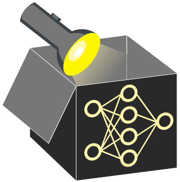
$$\mathbb{P}(X_j, X_k) = \mathbb{P}(X_j) \cdot \mathbb{P}(X_k)$$

- Equivalent definition (knowing X_k gives no info about X_j and vice versa):

  $\mathbb{P}(X_j | X_k) = \mathbb{P}(X_j)$ and $\mathbb{P}(X_k | X_j) = \mathbb{P}(X_k)$ (follows from cond. probability)

- Measuring complex dependencies is difficult but different measures exist Examples
  $\rightsquigarrow$ Spearman correlation (measures monotonic dependencies via ranks)
  $\rightsquigarrow$ Information-theoretical measures like mutual information
  $\rightsquigarrow$ Kernel-based measures like Hilbert-Schmidt Independence Criterion (HSIC)
- **N.B.:** X_j, X_k indep. $\Rightarrow \rho(X_j, X_k) = 0$ **but** $\rho(X_j, X_k) = 0 \not\Rightarrow X_j$, X_k indep. Equivalency holds if distribution is jointly normal

MUTUAL INFORMATION

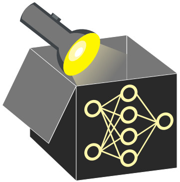


- MI describes expected amount of information shared by two RVs:

$$MI(X_1, X_2) = \mathbb{E}_{p(x_1,x_2)}\left[log\left(\frac{p(x_1, x_2)}{p(x_1)p(x_2)}\right)\right]$$

- MI measures amount of "dependence" between features by looking how different the joint distribution is from pure indep. $p(x_1, x_2) = p(x_1)p(x_2)$
  $\rightsquigarrow MI(X_1, X_2) = \mathbb{E}_{p(x_1,x_2)}\left[log\left(\frac{p(x_1,x_2)}{p(x_1,x_2)}\right)\right] = \mathbb{E}_{p(x_1,x_2)}\left[log(1)\right] = 0$
  $\rightsquigarrow MI(X_j, X_k) = 0$ if and only if the features are independent
- Unlike (Pearson) correlation, MI is also defined for categorical features

MUTUAL INFORMATION: EXAMPLE



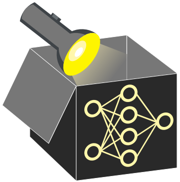
For two discrete RV X_1 and Y:

$$MI(X_1;Y) = \mathbb{E}_{p(x_1,y)}\left[log\left(\frac{p(x_1,y)}{p(x_1)p(y)}\right)\right] = \sum_{x_1\in\mathcal{X}_1}\sum_{y\in\mathcal{Y}} p(x_1,y) log\left(\frac{p(x_1,y)}{p(x_1)p(y)}\right)$$

X_1	...	**Y**
yes	...	yes
yes	...	no
no	...	yes
no	...	no

	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = \text{no})$	p_Y
$\mathbb{P}(Y = \text{yes})$	0.25	0.25	0.5
$\mathbb{P}(Y = \text{no})$	0.25	0.25	0.5
p_{X_1}	0.5	0.5	1

MUTUAL INFORMATION: EXAMPLE



For two discrete RV X_1 and Y:

$$MI(X_1;Y) = \mathbb{E}_{p(x_1,y)}\left[log\left(\frac{p(x_1,y)}{p(x_1)p(y)}\right)\right] = \sum_{x_1\in\mathcal{X}_1}\sum_{y\in\mathcal{Y}} p(x_1,y)log\left(\frac{p(x_1,y)}{p(x_1)p(y)}\right)$$

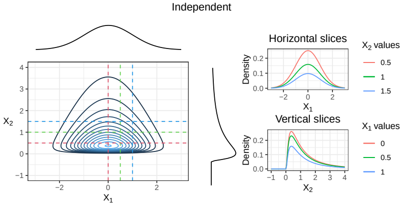
$\mathbf{X_1}$	...	**Y**
yes	...	yes
yes	...	no
no	...	yes
no	...	no

	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = \text{no})$	p_Y
$\mathbb{P}(Y = \text{yes})$	0.25	0.25	0.5
$\mathbb{P}(Y = \text{no})$	0.25	0.25	0.5
p_{X_1}	0.5	0.5	1

$$\begin{aligned} MI(X_1;Y) &= 0.25\,log\left(\frac{0.25}{0.5\cdot 0.5}\right) + 0.25\,log\left(\frac{0.25}{0.5\cdot 0.5}\right) \\ &\quad + 0.25\,log\left(\frac{0.25}{0.5\cdot 0.5}\right) + 0.25\,log\left(\frac{0.25}{0.5\cdot 0.5}\right) \\ &= 0.25\,log\left(\frac{0.25}{0.25}\right)\cdot 4 \\ &= 0.25\,log\left(1\right)\cdot 4 = 0 \end{aligned}$$

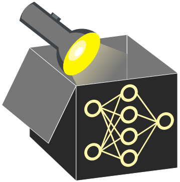
DEPENDENCE AND INDEPENDENCE

Example:



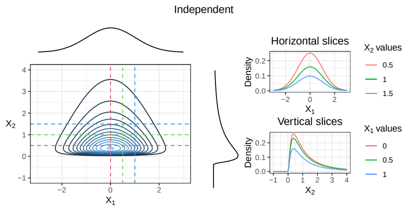
Conditional distributions at different vertical and horizontal slices (after normalizing area to 1) match their marginal distributions

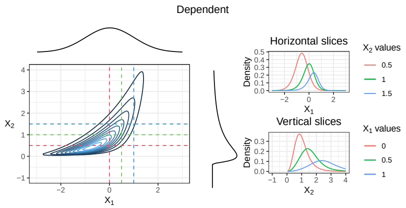
$$\Rightarrow \mathbb{P}(X_1|X_2) = \mathbb{P}(X_1)$$
$$\mathbb{P}(X_2|X_1) = \mathbb{P}(X_2)$$



DEPENDENCE AND INDEPENDENCE

Example:





Conditional distributions at different vertical and horizontal slices (after normalizing area to 1) match their marginal distributions

Conditional distributions do not match their marginal distributions

$$\Rightarrow \mathbb{P}(X_1|X_2) = \mathbb{P}(X_1)$$
$$\mathbb{P}(X_2|X_1) = \mathbb{P}(X_2)$$

CORRELATION VS. DEPENDENCE

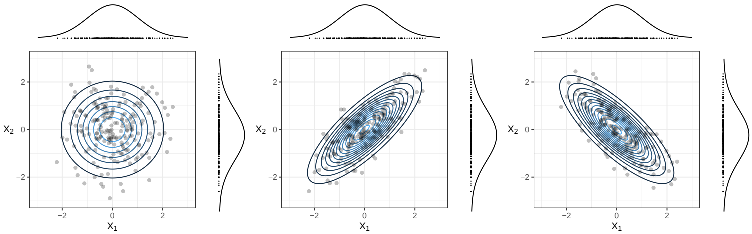
Illustration of bivariate normal distribution with different correlations

$X_1, X_2 \sim N(0, 1)$

$\rho(X_1, X_2) = 0$ (independent)

$\rho(X_1, X_2) = 0.8$

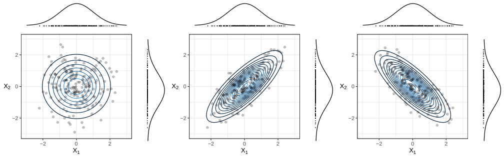
$\rho(X_1, X_2) = -0.8$



CORRELATION VS. DEPENDENCE

Illustration of bivariate normal distribution with different correlations $X_1, X_2 \sim N(0, 1)$

$\rho(X_1, X_2) = 0$ (independent) $\qquad \rho(X_1, X_2) = 0.8 \qquad \rho(X_1, X_2) = -0.8$



Examples with Pearson's corr. $\rho \approx 0$ but non-linear dependencies (MI $\neq 0$):

$\rho(X_1, X_2) = 0$, $MI(X_1, X_2) = 0.52 \qquad \rho(X_1, X_2) = 0.01$, $MI(X_1, X_2) = 0.37 \qquad \rho(X_1, X_2) = -0.06$, $MI(X_1, X_2) = 0.61$