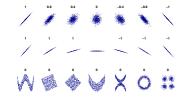
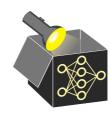
# **Interpretable Machine Learning**

# **Correlation and Dependencies**



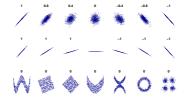
## Learning goals

- Pearson correlation
- Coefficient of determination R<sup>2</sup>
- Mutual information
- Correlation vs. dependence



# **Interpretable Machine Learning Correlation and Dependencies**





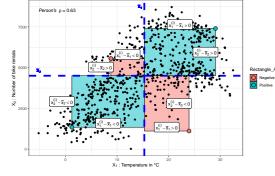
## Learning goals

- Pearson correlation
- Coefficient of determination R<sup>2</sup>
- Mutual information
- Correlation vs. dependence

# PEARSON'S CORRELATION COEFFICIENT $\rho$

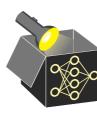
**Correlation** often refers to Pearson's correlation (measures only **linear relationship**)

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



Geometric interpretation of  $\rho$ :

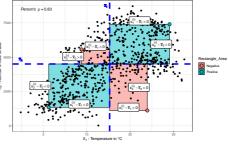
- Areas enter numerator with positive (+) or negative (-) sign, depending on position
- Denominator scales the sum into the range [-1, 1]



# PEARSON'S CORRELATION COEFFICIENT $\rho$

**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  $\rho$ :

- Areas enter numerator with positive (+) or negative (-) sign, depending on position
- Denominator scales the sum into the range [-1, 1]

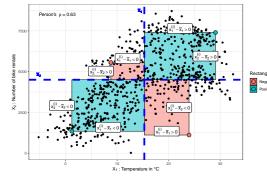


Interpretable Machine Learning - 1/8 © -1/8

# PEARSON'S CORRELATION COEFFICIENT $\rho$

**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  $\rho$ :

- Areas enter numerator with positive (+) or negative (-) sign, depending on position
- Denominator scales the sum into the range [-1, 1]

Interpretable Machine Learning - 1/8

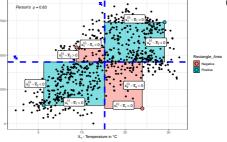
- $\rho > 0$  if positive areas dominate negative areas  $\rightsquigarrow X_1, X_2$  positive correlated
- $\rho$  < 0 if negative areas dominate positive areas  $\rightsquigarrow$   $X_1$ ,  $X_2$  negative correlated
- $\rho = 0$  if area of rectangles cancels out  $\rightsquigarrow X_1, X_2$  linearly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



# PEARSON'S CORRELATION COEFFICIENT $\rho$

**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  $\rho$ :

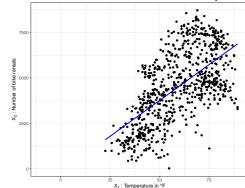
- Numerator is sum of rectangle's area with width x<sub>1</sub><sup>(i)</sup> − x̄<sub>1</sub> and height x<sub>2</sub><sup>(i)</sup> − x̄<sub>2</sub>
- Areas enter numerator with positive (+) or negative (-) sign, depending on position
- Denominator scales the sum into the range [-1, 1]
- $\rho > 0$  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



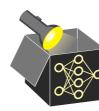
# COEFFICIENT OF DETERMINATION $R^2$

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



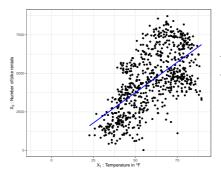


- Fit a linear model:
- $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$
- $\rightsquigarrow$  Slope  $\theta_1 = 0 \Rightarrow$  no dependence
- $\leadsto \ \text{Large slope} \Rightarrow \text{strong dependence}$



# COEFFICIENT OF DETERMINATION $R^2$

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

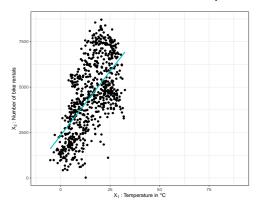


- Fit a linear model:
- $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$
- $\rightarrow$  Slope  $\theta_1 = 0 \Rightarrow$  no dependence
- $\leadsto$  Large slope  $\Rightarrow$  strong dependence



# COEFFICIENT OF DETERMINATION $R^2$

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



• Fit a linear model:

$$\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$$

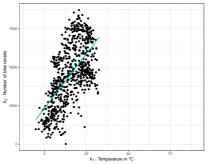
- $\rightsquigarrow$  Slope  $\theta_1 = 0 \Rightarrow$  no dependence
- $\leadsto$  Large slope  $\Rightarrow$  strong dependence
- Exact  $\theta_1$  score problematic
- $\rightarrow$  Re-scaling of  $x_1$  or  $x_2$  changes  $\theta_1$

$$ightharpoonup$$
°F  $ightharpoonup$ °C  $\Rightarrow \theta_1 = 78 \rightarrow \theta_1^* = 141$ 



# COEFFICIENT OF DETERMINATION $R^2$

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



• Fit a linear model:

$$\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$$

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

- $\rightsquigarrow \ \, \text{Large slope} \Rightarrow \text{strong dependence}$
- Exact  $\theta_1$  score problematic
- $\rightsquigarrow$  Re-scaling of  $x_1$  or  $x_2$  changes  $\theta_1$

$$ightsquigarrow \ {}^{\circ}\mathsf{F} 
ightarrow \ {}^{\circ}\mathsf{C} \Rightarrow \theta_1 = 78 
ightarrow \theta_1^* = 141$$

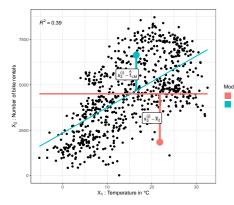


Interpretable Machine Learning - 2/8

© - 2/8

# COEFFICIENT OF DETERMINATION R<sup>2</sup>

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



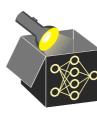
- Fit a linear model:  $\hat{x}_2 = \hat{f}_{LM}(x_1) = \theta_0 + \theta_1 x_1$
- $\rightarrow$  Slope  $\theta_1 = 0 \Rightarrow$  no dependence
- → Large slope ⇒ strong dependence
- Exact  $\theta_1$  score problematic
- $\rightarrow$  Re-scaling of  $x_1$  or  $x_2$  changes  $\theta_1$
- Set SSE<sub>IM</sub> in relation to SSE of a constant model  $\hat{f}_c = \bar{x}_2$  $SSE_{LM} = \sum_{i=1}^{n} (x_2^{(i)} - \hat{f}_{LM}(x_1^{(i)}))^2$

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

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

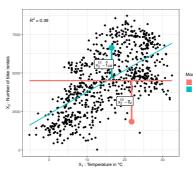
$$\Rightarrow$$
 Measure of fitting quality of LM:  $R^2 = 1 - \frac{SSE_{LM}}{SSE_o} \in [0, 1]$ 

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



# COEFFICIENT OF DETERMINATION R<sup>2</sup>

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
- → Large slope ⇒ strong dependence
- Exact  $\theta_1$  score problematic
- $\rightarrow$  Re-scaling of  $x_1$  or  $x_2$  changes  $\theta_1$
- Set SSE<sub>LM</sub> 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:  $R^2 = 1 - \frac{SSE_{LM}}{SSE} \in [0, 1]$ 

$$\Rightarrow$$
 Measure of fitting quality of LM:  $R^2 = \Rightarrow \rho(X_1, X_2) = R$ 



Interpretable Machine Learning - 2 / 8 - 2/8

# JOINT, MARGINAL AND CONDITIONAL DISTRIBUTION

 $p_{X_1,X_2}$ 

 $\mathbb{P}(X_1 = 0)$ 

 $\mathbb{P}(X_1=1)$ 

 $\mathbb{P}(X_2=0) \quad \mathbb{P}(X_2=1) \quad p_{X_1}$ 

0.4

0.7

0.5

0.2

0.3

For two discrete random variables  $X_1, X_2$ :

## Joint distribution

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



# JOINT, MARGINAL AND CONDITIONAL DISTRIBUTION

For two discrete random variables  $X_1, X_2$ :

## Joint distribution

Joint distribution	$p_{X_1,X_2}$	$\mathbb{P}(X_2=0)$	$\mathbb{P}(X_2=1)$	$p_{X_1}$
	$\mathbb{P}(X_1=0)$	0.2	0.3	0.5
- (v v) TD(V v V v)	$\mathbb{P}(X_1=1)$	0.1	0.4	0.5
$p_{X_1,X_2}(x_1,x_2) = \mathbb{P}(X_1 = x_1,X_2 = x_2)$	$p_{X_2}$	0.3	0.7	1



Interpretable Machine Learning - 3/8 - 3/8

# 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**

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

$p_{X_1,X_2}$	$\mathbb{P}(X_2=0)$	$\mathbb{P}(X_2=1)$	$p_{X_1}$
$\mathbb{P}(X_1=0)$	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

→ 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
$p_{X_0}$	0.3	0.7	1



# Marginal distribution

$$p_{X_1}(x_1) = \mathbb{P}(X_1 = x_1) = \sum_{x_2 \in \mathcal{X}_2} p(x_1, x_2) \begin{vmatrix} p_{X_1, x_2} & \mathbb{P}(X_2 = 0) & \mathbb{P}(X_2 = 1) & p_{X_1} \\ \mathbb{P}(X_1 = 0) & 0.2 & 0.3 & 0.5 \\ \mathbb{P}(X_1 = 1) & 0.1 & 0.4 & 0.5 \\ \mathbb{P}(X_1 = 0) & 0.3 & 0.7 & 1 \\ \mathbb{P}(X_1 = 0) & 0.7 & 0.7 & 1 \\ \mathbb{P}(X_1 = 0) & 0.7 & 0.7 & 1 \\ \mathbb{P}(X_1 = 0) & 0.7 & 0.7 & 1 \\ \mathbb{P}(X_1 = 0) & 0.7 & 0.7 & 1 \\ \mathbb{P}(X_1 = 0) & 0.7 & 0.7 & 1 \\ \mathbb{P}(X_1 = 0) & 0.7 & 0.7 & 1 \\ \mathbb{P}(X_1 = 0) & 0.7 & 0.7 & 1 \\ \mathbb{P}(X_1 = 0) & 0.7 & 0.7 & 1 \\ \mathbb{P}(X_1 = 0) & 0.7 & 0.7 & 1 \\ \mathbb{P}(X_1 = 0) & 0.7 & 0.7 & 1 \\ \mathbb{P}(X_1 = 0) & 0.7 & 0.7 & 1 \\ \mathbb{P}(X_1 = 0) & 0.7 & 0.7 & 1 \\$$

→ In continuous case with integrals

Interpretable Machine Learning - 3/8

# JOINT, MARGINAL AND CONDITIONAL DISTRIBUTION

 $p_{X_1,X_2}$ 

 $\mathbb{P}(X_1 = 0)$ 

 $\mathbb{P}(X_1=1)$ 

For two discrete random variables  $X_1, X_2$ :

### Joint distribution

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

1)	$p_{X_1}$	
	0.5	
	0.5	
	1	

# **Marginal distribution**

$$p_{X_1}(x_1) = \mathbb{P}(X_1 = x_1) = \sum_{i=1}^{n} p(x_1, x_2)$$

$p_{X_1,X_2}$	$\mathbb{P}(X_2=0)$	$\mathbb{P}(X_2=1)$	$p_{X_1}$
$\mathbb{P}(X_1=0)$	0.2	0.3	0.5
$\mathbb{P}(X_1=1)$	0.1	0.4	0.5
$p_{X_0}$	0.3	0.7	1

 $\mathbb{P}(X_2=0) \mid \mathbb{P}(X_2=$ 

0.2

0.3

→ 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

# JOINT. MARGINAL AND CONDITIONAL DISTRIBUTION

For two discrete random variables  $X_1, X_2$ :

### Joint distribution

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



# Marginal distribution

$$\rho_{X_1}(x_1) = \mathbb{P}(X_1 = x_1) = \sum_{x_2 \in \mathcal{X}_2} \rho(x_1, x_2) \begin{vmatrix} \rho_{X_1, X_2} & \mathbb{P}(X_2 = 0) & \mathbb{P}(X_2 = 1) & \rho_{X_1} \\ \mathbb{P}(X_1 = 0) & 0.2 & 0.3 & 0.5 \\ \mathbb{P}(X_1 = 1) & 0.1 & 0.4 & 0.5 \end{vmatrix}$$

→ In continuous case with integrals

# **Conditional distribution**

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

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

Interpretable Machine Learning - 3/8 - 3/8

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

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

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

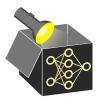


# **DEPENDENCE**

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

• Definition:  $X_i$ ,  $X_k$  independent  $\Leftrightarrow$  joint distribution is product of marginals:

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



**Dependence:** 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 (knowledge of  $X_k$  says nothing about  $X_i$  and vice versa):

$$\mathbb{P}(X_i|X_k) = \mathbb{P}(X_i)$$
 and  $\mathbb{P}(X_k|X_i) = \mathbb{P}(X_k)$  (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  $X_i$  and vice versa):

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



Interpretable Machine Learning - 4 / 8 © - 4/8

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

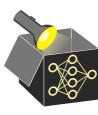
• Definition:  $X_i$ ,  $X_k$  independent  $\Leftrightarrow$  joint distribution is product of marginals:

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

• Equivalent definition (knowledge of  $X_k$  says nothing about  $X_i$  and vice versa):

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

- Measuring complex dependencies is difficult but different measures exist, e.g.,
  - → Spearman correlation (measures monotonic dependencies via ranks)
  - → Information-theoretical measures like mutual information



## DEPENDENCE

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



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

• Equivalent definition (knowing  $X_k$  gives no info about  $X_i$  and vice versa):

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

- Measuring complex dependencies is difficult but different measures exist Examples
- Spearman correlation (measures monotonic dependencies via ranks)
- → Information-theoretical measures like mutual information



Interpretable Machine Learning - 4/8 © - 4/8

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

• Definition:  $X_i$ ,  $X_k$  independent  $\Leftrightarrow$  joint distribution is product of marginals:

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

• Equivalent definition (knowledge of  $X_k$  says nothing about  $X_i$  and vice versa):

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

- Measuring complex dependencies is difficult but different measures exist, e.g.,
  - → Spearman correlation (measures monotonic dependencies via ranks)
  - → Information-theoretical measures like mutual information
  - → Kernel-based measures like Hilbert-Schmidt Independence Criterion (HSIC)
- **N.B.:**  $X_j$ ,  $X_k$  independent  $\Rightarrow \rho(X_j, X_k) = 0$  **but**  $\rho(X_j, X_k) = 0 \Rightarrow X_j$ ,  $X_k$  indep. Equivalency holds if distribution is jointly normal



## DEPENDENCE

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

• Definition:  $X_i$ ,  $X_k$  independent  $\Leftrightarrow$  joint distribution is product of marginals:

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

• Equivalent definition (knowing  $X_k$  gives no info about  $X_i$  and vice versa):

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

- Measuring complex dependencies is difficult but different measures exist Examples
- → Spearman correlation (measures monotonic dependencies via ranks)
- → Information-theoretical measures like mutual information
   → Kernel-based measures like Hilbert-Schmidt Independence Criterion (HSIC)
- **N.B.:**  $X_j$ ,  $X_k$  indep.  $\Rightarrow \rho(X_j, X_k) = 0$  **but**  $\rho(X_j, X_k) = 0 \Rightarrow X_j$ ,  $X_k$  indep. Equivalency holds if distribution is jointly normal



Interpretable Machine Learning - 4/8 © -4/8

# **MUTUAL INFORMATION**

MI describes expected amount of information shared by two random variables:

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

 MI measures amount of "dependence" between features by looking how different the joint distribution is from pure independence  $p(x_1, x_2) = p(x_1)p(x_2)$  $ightharpoonup MI(X_1, X_2) = \mathbb{E}_{p(x_1, x_2)} \left[ log \left( \frac{p(x_1, x_2)}{p(x_1, x_2)} \right) \right] = \mathbb{E}_{p(x_1, x_2)} \left[ log(1) \right] = 0$ 

$$\rightsquigarrow MI(X_j, X_k) = 0$$
 if and only if the features are independent

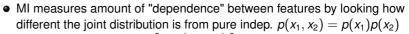
Unlike (Pearson) correlation, MI can also be computed for categorical features



## MUTUAL INFORMATION

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

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



$$\rightsquigarrow MI(X_1, X_2) = \mathbb{E}_{\rho(X_1, X_2)} \left[ log \left( \frac{\rho(X_1, X_2)}{\rho(X_1, X_2)} \right) \right] = \mathbb{E}_{\rho(X_1, X_2)} \left[ log(1) \right] = 0$$

 $\rightsquigarrow MI(X_i, X_k) = 0$  if and only if the features are independent

• Unlike (Pearson) correlation, MI is also defined for categorical features



Interpretable Machine Learning - 5 / 8 - 5/8

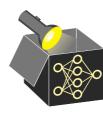
# MUTUAL INFORMATION: EXAMPLE

For two discrete RV  $X_1$  and Y:

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

X <sub>1</sub>	 Y
yes	 yes
yes	 no
no	 yes
no	 no

	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = no)$	p <sub>Y</sub>
$\mathbb{P}(Y = \text{yes})$	0.25	0.25	0.5
$\mathbb{P}(Y = no)$	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:

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



1	
yes	 yes
yes	 no
no	 yes
no	 no

	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = no)$	p <sub>Y</sub>
$\mathbb{P}(Y = \text{yes})$	0.25	0.25	0.5
$\mathbb{P}(Y = no)$	0.25	0.25	0.5
$p_{X_1}$	0.5	0.5	1

Interpretable Machine Learning - 6/8

# **MUTUAL INFORMATION: EXAMPLE**

For two discrete RV  $X_1$  and Y:

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

<b>X</b> <sub>1</sub>	 Υ
yes	 yes
yes	 no
no	 yes
no	 no

	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = no)$	p <sub>Y</sub>
$\mathbb{P}(Y = \text{yes})$	0.25	0.25	0.5
$\mathbb{P}(Y = no)$	0.25	0.25	0.5
$p_{X_1}$	0.5	0.5	1

$$MI(X_1; Y) = 0.25 \log \left(\frac{0.25}{0.5 \cdot 0.5}\right) + 0.25 \log \left(\frac{0.25}{0.5 \cdot 0.5}\right)$$

$$= 0.25 \log \left(\frac{0.25}{0.25}\right) \cdot 4$$

$$= 0.25 \log (1) \cdot 4 = 0$$

# **MUTUAL INFORMATION: EXAMPLE**

For two discrete RV  $X_1$  and Y:

For two discrete RV 
$$X_1$$
 and  $Y$ 

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



- 6/8

1		
yes	 yes	
yes	 no	
no	 yes	
no	 no	

	$\mathbb{P}(X_1 = \text{yes})$	$\mathbb{P}(X_1 = no)$	p <sub>Y</sub>
$\mathbb{P}(Y = \text{yes})$ $\mathbb{P}(Y = \text{no})$	0.25	0.25	0.5
$\mathbb{P}(Y = no)$	0.25	0.25	0.5
$p_{X_1}$	0.5	0.5	1

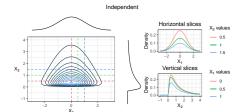
$$MI(X_1; Y) = 0.25 \log \left(\frac{0.25}{0.5 \cdot 0.5}\right) + 0.25 \log \left(\frac{0.25}{0.5 \cdot 0.5}\right)$$

$$= 0.25 \log \left(\frac{0.25}{0.25}\right) \cdot 4$$

$$= 0.25 \log (1) \cdot 4 = 0$$

# DEPENDENCE AND INDEPENDENCE

# Example:



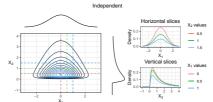
Conditional distributions at different vertical and horizontal slices (after normalizing area to 1) match their marginal distributions

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



# DEPENDENCE AND INDEPENDENCE

# Example:



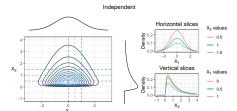


Conditional distributions at different vertical and horizontal slices (after normalizing area to 1) match their marginal distributions

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

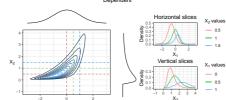
# DEPENDENCE AND INDEPENDENCE

## Example:



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

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

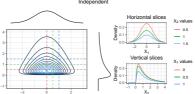


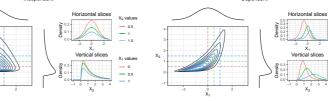
Conditional distributions do not match their marginal distributions



# **DEPENDENCE AND INDEPENDENCE**

## Example:





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

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

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

- 7/8

Conditional distributions do not match their marginal distributions

# **CORRELATION VS. DEPENDENCE**

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

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



# CORRELATION VS. DEPENDENCE

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

$$ho(X_1, X_2) = 0$$
  $ho(X_1, X_2) = 0.8$   $ho(X_1, X_2) = -0.8$  (independent)



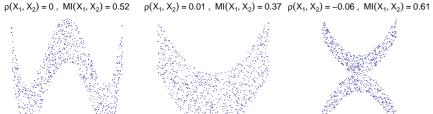
# **CORRELATION VS. DEPENDENCE**

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

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

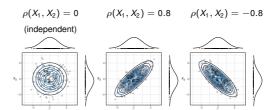


Examples with Pearson's correlation  $\rho \approx 0$  but non-linear dependencies (MI  $\neq 0$ ):



# **CORRELATION VS. DEPENDENCE**

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





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



