CMT311 Principles of Machine Learning

Generative Models

Angelika Kimmig
KimmigA@cardiff.ac.uk

08.11.2019

ERM Learning

with randomly labeled examples

the set of objects we want to label

domain set X

{0,1}

label set Y

the possible hypotheses

hypothesis space H

data-generation model → **training data S** → **ERM learner**

a finite sequence of objects from X labeled with elements of Y

a probability distribution D over $X \times Y$

hypothesis $h_S \in \arg\min_{h \in H} L_S(h)$

the learner does not know D

i.i.d. assumption, $S \sim D^m$: S contains m examples that are independently and identically distributed according to D

The Bayes optimal predictor

- For any D over $X \times \{0,1\}$, the best labeling function is
$$f_D(x) = \begin{cases} 1 & \text{if } P_D(y = 1 \mid x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

- best = no other $g : X \to \{0,1\}$ has lower true error

- but we do not know D …

- instead, we'll aim to learn a predictor whose error is not much larger than the best error in a given class of predictors

A different view

- So far: focus on **labelling**
  - assume the labelling function takes a specific form, and
  - learn a **discriminative model**, i.e., a labelling function of that form that performs well across all examples
- Alternative: focus on **data generation**
  - assume the unknown distribution takes a specific form, and
  - learn a **generative model**, i.e., a distribution of that form that is close to the true distribution

Generative Models

- Generative models are useful beyond Boolean concept learning
  - general way to capture uncertainty
  - not tied to a single task
- e.g., diagnosis: rules (or logic) fail for several reasons
  - laziness
  - theoretical ignorance
  - practical ignorance

Probability Theory

- **Probability theory**
  - provides a tool for dealing with **degrees of belief**
  - lets us summarise the **uncertainty** coming from laziness and ignorance
  - makes statements about **knowledge states** rather than "the world as it really is"
- We will mostly focus on the discrete (countable) case here

Describing possible situations

- **Random variables** X_i with associated **domains** dom(X_i)
- A **basic event** sets a random variable (RV) to an element of its domain, i.e., for some i, $X_i = e_i$ where $e_i \in dom(X_i)$
- A **possible world** $\boldsymbol{\omega}$ contains a basic event for each RV, i.e., $\omega = (X_1 = e_1, \ldots, X_n = e_n)$ with $e_i \in dom(X_i)$ for all i, sometimes also written $\omega = (e_1, \ldots, e_n)$
- **Sample space** Ω = set of all possible worlds = $dom(X_1) \times \cdots \times dom(X_n)$
- **Event** = basic event or a (nested) propositional formula over basic events (using $\neg$, $\vee$, $\wedge$) = a **set** of possible worlds

Probability Distributions

- A **probability distribution** is a function $P : \Omega \to \mathbb{R}$ such that
  - $0 \leq P(\omega) \leq 1$ for every $\omega \in \Omega$
  - $\sum_{\omega \in \Omega} P(\omega) = 1$
- also called **joint distribution** and written $P(X_1, \ldots, X_n)$
- sufficient to obtain the probability of any event E:
  $P(E) = \sum_{\omega \in E} P(\omega)$

Example

- Three countries (England, Scotland, Wales) and three (first) languages (English, Scottish, Welsh)
- dom(C) = {E,S,W}, dom(L) = {Eng,Scot,Wel}
- (made up) joint distribution P(C,L):

P(C,L)	C=E	C=S	C=W
L=Eng	0.836	0.056	0.024
L=Scot	0.0352	0.024	0
L=Wel	0.0088	0	0.016

Marginalisation

- Given joint distribution P(X,Y), the **marginal distribution** of X is defined by $P(X) = \sum_{y \in dom(Y)} P(X, y)$

- More generally:
$$P(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) = \sum_{x_i \in dom(X_i)} P(X_1, \ldots, X_{i-1}, x_i, X_{i+1}, \ldots, X_n)$$

- also called **summing out**

Example

P(C,L)	C=E	C=S	C=W
L=Eng	0.836	0.056	0.024
L=Scot	0.0352	0.024	0
L=Wel	0.0088	0	0.016

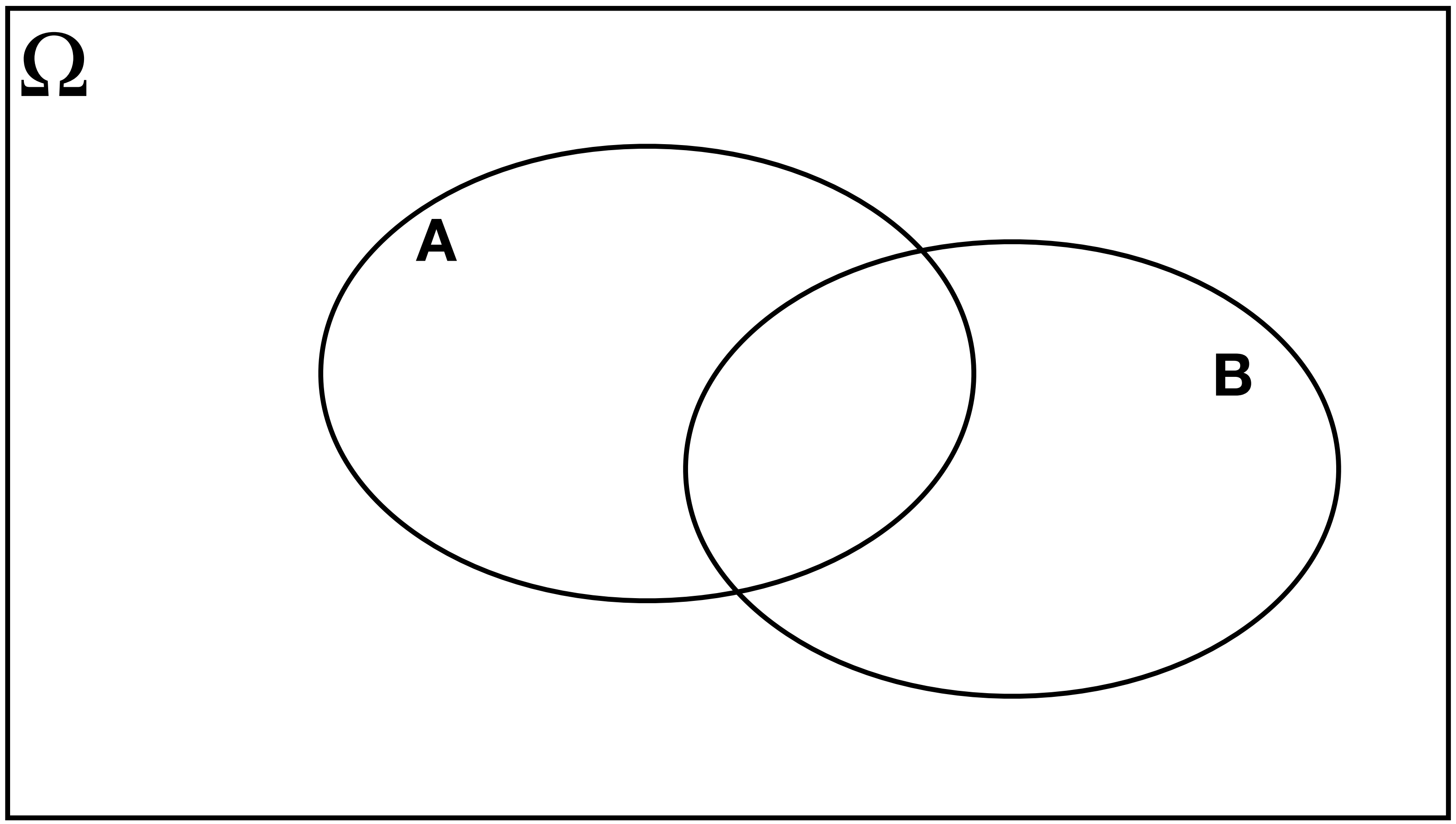
summing out C gives marginal P(L):

L=Eng	0.916
L=Scot	0.0592
L=Wel	0.0248

summing out L gives marginal P(C):

C=E	C=S	C=W
0.88	0.08	0.04

Probabilities of Events



$$P(\emptyset) = 0$$
$$P(\Omega) = 1$$
$$P(A \wedge B) = P(A, B) = P(A \cap B)$$
$$P(\neg A) = 1 - P(A)$$
$$P(A \vee B) = P(A) + P(B) - P(A \wedge B) = P(A \cup B)$$

Conditional Probability

- The **conditional probability** of event A **given** knowledge of event B is defined as $P(A \mid B) = \dfrac{P(A \wedge B)}{P(B)}$ if $P(B) > 0$ (and undefined otherwise)

- B is also called **evidence**

- **product rule:** $P(A \wedge B) = P(A \mid B) \cdot P(B)$

- **Bayes' rule:** $P(A \mid B) = \dfrac{P(B \mid A) \cdot P(A)}{P(B)}$

Example

P(C,L)	C=E	C=S	C=W
L=Eng	0.836	0.056	0.024
L=Scot	0.0352	0.024	0
L=Wel	0.0088	0	0.016

$$P(C = W \mid L = Wel) = \frac{P(C = W \wedge L = Wel)}{P(L = Wel)} = \frac{0.016}{0.0088 + 0 + 0.016} = 0.645$$

$$P(L = Eng \mid C = S \vee C = W) = \frac{P(L = Eng \wedge (C = S \vee C = W))}{P(C = S \vee C = W)}$$

$$= \frac{0.056 + 0.024}{0.056 + 0.024 + 0.024 + 0 + 0 + 0.016} = 0.667$$

Example

Prior probability P(C) based on population per country:

C=E	C=S	C=W
0.88	0.08	0.04

Conditional probability P(L|C) based on research:

P(L|C)	C=E	C=S	C=W
L=Eng	0.95	0.7	0.6
L=Scot	0.04	0.3	0
L=Wel	0.01	0	0.4

Product rule gives joint distribution:

P(C,L)	C=E	C=S	C=W
L=Eng	0.95x0.88	0.7x0.08	0.6x0.04
L=Scot	0.04x0.88	0.3x0.08	0x0.04
L=Wel	0.01x0.88	0x0.08	0.4x0.04

	C=E	C=S	C=W
P(C)	0.88	0.08	0.04

P(L|C)	C=E	C=S	C=W
L=Eng	0.95	0.7	0.6
L=Scot	0.04	0.3	0
L=Wel	0.01	0	0.4

What is P(C | L=Eng)?

Bayes' rule:

$$P(C \mid L = Eng) = \frac{P(L = Eng \mid C) \cdot P(C)}{P(L = Eng)}$$

P(C|L=Eng)	C=E	C=S	C=W
L=Eng	(0.95*0.88)/ P(L=Eng) = 0.836/P(L=Eng)	(0.7*0.08)/ P(L=Eng) = 0.056/P(L=Eng)	(0.6*0.04)/ P(L=Eng) = 0.024/P(L=Eng)

$$1 = \frac{0.836}{P(L = Eng)} + \frac{0.056}{P(L = Eng)} + \frac{0.024}{P(L = Eng)} = \frac{1}{P(L = Eng)}(0.836 + 0.056 + 0.024)$$

P(C|L=Eng)	C=E	C=S	C=W
L=Eng	0.9127	0.0611	0.0262

Independence

- Random variables X and Y are **independent**, written X⫫Y, if knowing the state of one variable gives no extra information about the other variable: $P(X, Y) = P(X) \cdot P(Y)$
- alternatively: $P(X \mid Y) = P(X)$ (and $P(Y) = P(Y \mid X)$)

Example

Joint distribution of the weather (W) and winning a bet (B)

P(W,B)	W=rain	W=sun
B=win	0.0175	0.0325
B=loss	0.3325	0.6175

marginals

P(B)

B=win	0.05
B=loss	0.95

P(W)

W=rain	W=sun
0.35	0.65

P(W)*P(B)	W=rain	W=sun
B=win	0.35*0.05=0.0175	0.65*0.05=0.0325
B=loss	0.35*0.95=0.3325	0.65*0.95=0.6175

W and B are independent

Homework: compute P(B|W) and P(W|B) and verify that the alternative characterisations on the previous slide indeed hold.

Conditional Independence

- Random variables X and Y are **conditionally independent** of each other **given** the state of random variable Z, written X⫫Y|Z, if $P(X, Y \mid Z) = P(X \mid Z) \cdot P(Y \mid Z)$

- Given the state of Z, knowing the state of X does not provide extra information about the state of Y (and vice versa)

- Also applies to **sets** of random variables: $\mathscr{X} \perp\!\!\!\perp \mathscr{Y} \mid \mathscr{Z}$ if $P(\mathscr{X}, \mathscr{Y} \mid \mathscr{Z}) = P(\mathscr{X} \mid \mathscr{Z}) \cdot P(\mathscr{Y} \mid \mathscr{Z})$ for all states of the variables in $\mathscr{X}, \mathscr{Y}, \mathscr{Z}$; we write $\mathscr{X} \perp\!\!\!\perp \mathscr{Y}$ for $\mathscr{X} \perp\!\!\!\perp \mathscr{Y} \mid \varnothing$

Example

Boolean variables **C**loudy, **S**prinkler and **R**ain

Claim: S⫫R|C

need to show:

$$P(S,R\,|\,C) = P(S\,|\,C)\cdot P(R\,|\,C)$$

$$P(S,R\,|\,C) = \frac{P(S,R,C)}{P(C)} \qquad P(S\,|\,C) = \frac{P(S,C)}{P(C)} \qquad P(R\,|\,C) = \frac{P(R,C)}{P(C)}$$

	R=yes		R=no	
P(S,R,C)	**S=yes**	**S=no**	**S=yes**	**S=no**
C=yes	0.04	0.36	0.01	0.09
C=no	0.05	0.05	0.20	0.20

P(C)

C=yes	0.04+0.36+0.01+0.09=0.5
C=no	0.05+0.05+0.2+0.2=0.5

P(S,C)	S=yes	S=no
C=yes	0.05	0.45
C=no	0.25	0.25

P(R,C)	R=yes	R=no
C=yes	0.4	0.1
C=no	0.1	0.4

	R=yes		R=no	
P(S,R|C)	**S=yes**	**S=no**	**S=yes**	**S=no**
C=yes	0.08	0.72	0.02	0.18
C=no	0.1	0.1	0.4	0.4

P(S|C)	S=yes	S=no
C=yes	0.1	0.9
C=no	0.5	0.5

P(R|C)	R=yes	R=no
C=yes	0.8	0.2
C=no	0.2	0.8

	R=yes		R=no	
P(S|C)*P(R|C)	**S=yes**	**S=no**	**S=yes**	**S=no**
C=yes	0.1*0.8	0.9*0.8	0.1*0.2	0.9*0.2
C=no	0.5*0.2	0.5*0.2	0.5*0.8	0.5*0.8

Example

Boolean variables **C**loudy, **S**prinkler and **R**ain

Note: S⫫R does not hold, i.e.,

$$P(S) \cdot P(R) \neq P(S, R)$$

	R=yes		R=no	
P(S,R,C)	S=yes	S=no	S=yes	S=no
C=yes	0.04	0.36	0.01	0.09
C=no	0.05	0.05	0.20	0.20

P(S)

S=yes	S=no
0.3	0.7

P(R)

R=yes	R=no
0.5	0.5

P(S,R)

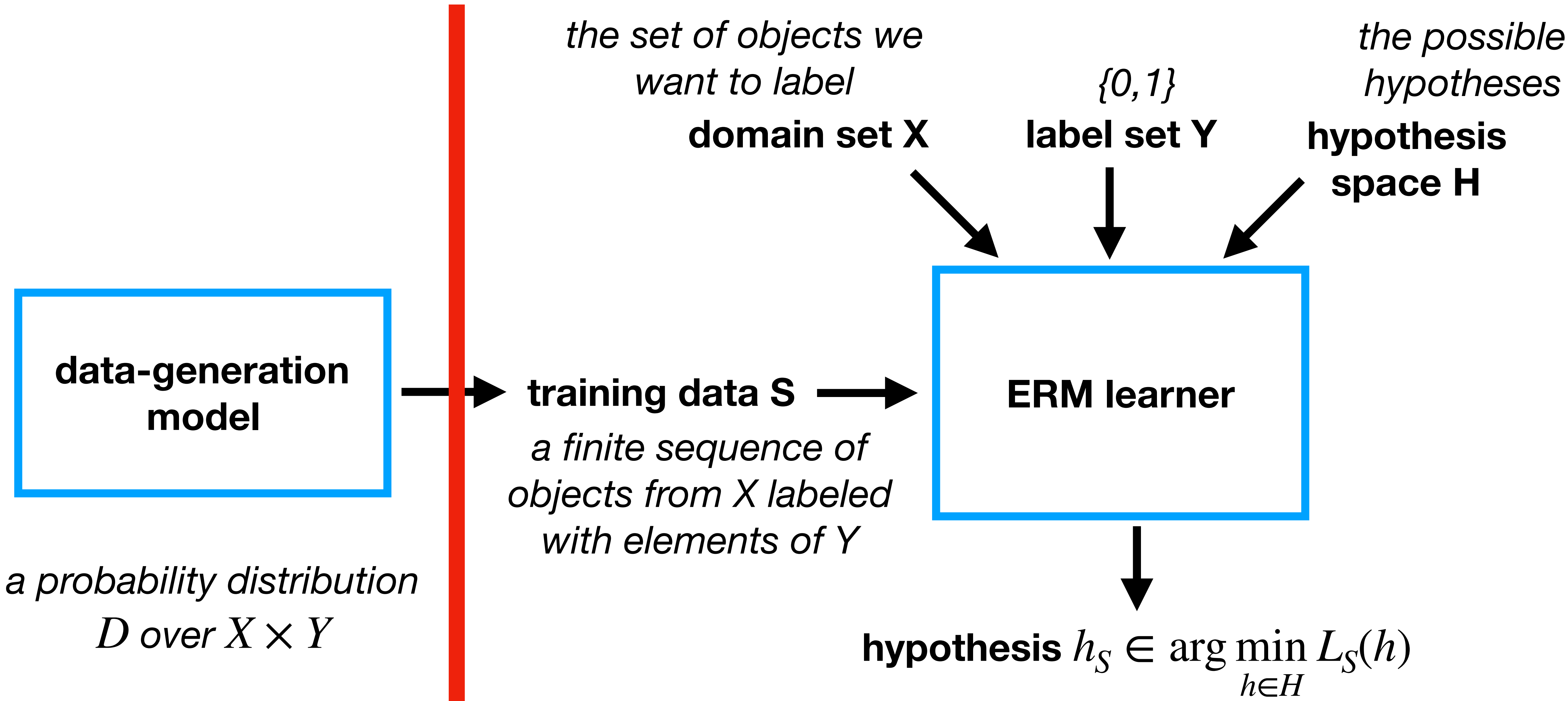
R=yes		R=no	
S=yes	S=no	S=yes	S=no
0.09	0.41	0.21	0.29

P(S)*P(R)

R=yes		R=no	
S=yes	S=no	S=yes	S=no
0.5*0.3=0.15	0.5*0.7=0.35	0.5*0.3=0.15	0.5*0.7=0.35

ERM Learning

with randomly labeled examples



the learner does not know D

i.i.d. assumption, $S \sim D^m$: S contains m examples that are independently and identically distributed according to D *days described by*

< Sky, AirTemp, Humidity, Wind, Water, Forecast > + label enjoy

Sunny / Cloudy / Rainy

Warm / Cold

Normal / High

Strong / Weak

Warm / Cool

Same / Change

0 / 1

$$\Omega = \{\text{Sunny,Cloudy,Rainy}\} \text{x} \{\text{Warm,Cold}\} \text{x} \{\text{Normal,High}\} \text{x} \{\text{Strong,Weak}\} \text{x} \{\text{Warm,Cool}\} \text{x} \{\text{Same,Change}\} \text{x} \{0,1\}$$

Sky	AirTemp	Humidity	Wind	Water	Forecast	Enjoy	P(ω)
Sunny	Warm	Normal	Strong	Warm	Same	0	p1
Sunny	Warm	Normal	Strong	Warm	Same	1	p2
Sunny	Warm	Normal	Strong	Warm	Change	0	p3
Sunny	Warm	Normal	Strong	Warm	Change	1	p4
Sunny	Warm	Normal	Strong	Cool	Same	0	p5
Sunny	Warm	Normal	Strong	Cool	Same	1	p6
Sunny	Warm	Normal	Strong	Cool	Change	0	p7
Sunny	Warm	Normal	Strong	Cool	Change	1	p8
Sunny	Warm	Normal	Weak	Warm	Same	0	p9
			…				
			…				
Rainy	Cold	High	Weak	Cool	Change	1	p192

all in [0,1],
sum = 1

Learning

- Choose:
  - a representation of a probability distribution over $X \times Y$ with parameters θ
  - a prior $P(\theta)$ over the values of the parameters
  - a generative model $P(S \mid \theta)$ for the data given the parameters
- Bayes' rule: $P(\theta \mid S) = \dfrac{P(S \mid \theta)P(\theta)}{P(S)}$
- The **MAP (most probable a posteriori) parameter estimate** is the one maximising the posterior, $\theta^{MAP} = \arg\max_{\theta} P(\theta \mid S) = \arg\max_{\theta} \dfrac{P(S \mid \theta)P(\theta)}{P(S)}$
- If $P(\theta)$ is equal for all values, the MAP estimate becomes the **ML (maximum likelihood) estimate**, $\theta^{ML} = \arg\max_{\theta} P(S \mid \theta)$

$$\theta^{ML} = \arg\max_{\theta} P(S \mid \theta)$$

remember $S \sim D^m$, i.e., each example in S is some row in the table, and $P(S \mid \theta)$ is the product of the corresponding parameters

let c_i be the number of times the i-th row appears in S

then, $\theta^{ML} = (\frac{c_1}{m}, \ldots, \frac{c_m}{m})$

Learning the parameters for the full joint distribution is unrealistic…

Sky	AirTemp	Humidity	Wind	Water	Forecast	Enjoy	P(ω)
Sunny	Warm	Normal	Strong	Warm	Same	0	p1
Sunny	Warm	Normal	Strong	Warm	Same	1	p2
Sunny	Warm	Normal	Strong	Warm	Change	0	p3
Sunny	Warm	Normal	Strong	Warm	Change	1	p4
Sunny	Warm	Normal	Strong	Cool	Same	0	p5
Sunny	Warm	Normal	Strong	Cool	Same	1	p6
Sunny	Warm	Normal	Strong	Cool	Change	0	p7
Sunny	Warm	Normal	Strong	Cool	Change	1	p8
Sunny	Warm	Normal	Weak	Warm	Same	0	p9
			…				
			…				
Rainy	Cold	High	Weak	Cool	Change	1	p192

all in [0,1], sum = 1

$\theta = (p_1, \ldots, p_{192})$

Learning

Impose some **structure** on the distribution using (conditional) independence

- Choose:
  - a representation of a probability distribution over $X \times Y$ with parameters θ
  - a prior $P(\theta)$ over the values of the parameters
  - a generative model $P(S \mid \theta)$ for the data given the parameters
- Bayes' rule: $P(\theta \mid S) = \dfrac{P(S \mid \theta)P(\theta)}{P(S)}$
- The **MAP (most probable a posteriori) parameter estimate** is the one maximising the posterior, $\theta^{MAP} = \arg\max_{\theta} P(\theta \mid S) = \arg\max_{\theta} \dfrac{P(S \mid \theta)P(\theta)}{P(S)}$
- If $P(\theta)$ is equal for all values, the MAP estimate becomes the **ML (maximum likelihood) estimate**, $\theta^{ML} = \arg\max_{\theta} P(S \mid \theta)$

days described by

< Sky, AirTemp, Humidity, Wind, Water, Forecast > + label enjoy

- Sky: *Sunny / Cloudy / Rainy*
- AirTemp: *Warm / Cold*
- Humidity: *Normal / High*
- Wind: *Strong / Weak*
- Water: *Warm / Cool*
- Forecast: *Same / Change*
- label enjoy: *0 / 1*

Let's assume the attributes are independent given the label:

$$P(S, A, H, Wi, Wa, F, E) = P(S \mid E)P(A \mid E)P(H \mid E)P(Wi \mid E)P(Wa \mid E)P(F \mid E)P(E)$$

E=0	E=1
p_0	p_1

P(S|E)	S=Sunny	S=Cloudy	S=Rainy
E=0	p_2	p_3	p_4
E=1	p_5	p_6	p_7

P(A|E)	A=Warm	A=Cold
E=0	p_8	p_9
E=1	p_{10}	p_{11}

P(H|E)	H=Normal	H=High
E=0	p_{12}	p_{13}
E=1	p_{14}	p_{15}

P(Wi|E)	Wi=Strong	Wi=Weak
E=0	p_{16}	p_{17}
E=1	p_{18}	p_{19}

P(Wa|E)	Wa=Warm	Wa=Cool
E=0	p_{20}	p_{21}
E=1	p_{22}	p_{23}

P(F|E)	F=Same	F=Change
E=0	p_{24}	p_{25}
E=1	p_{26}	p_{27}

Use Bayes' rule to determine the most likely label of a new example $< v_1, \ldots, v_6 >$:

$$\arg\max_{e \in \{0,1\}} P(E = e \,|\, S = v_1, A = v_2, H = v_3, Wi = v_4, Wa = v_5, F = v_6)$$

$$= \arg\max_{e \in \{0,1\}} \frac{P(S = v_1, A = v_2, H = v_3, Wi = v_4, Wa = v_5, F = v_6 \,|\, E = e)P(E = e)}{P(S = v_1, A = v_2, H = v_3, Wi = v_4, Wa = v_5, F = v_6)}$$

$$= \arg\max_{e \in \{0,1\}} P(S = v_1, A = v_2, H = v_3, Wi = v_4, Wa = v_5, F = v_6 \,|\, E = e)P(E = e)$$

$$= \arg\max_{e \in \{0,1\}} P(S = v_1 \,|\, E = e)P(A = v_2 \,|\, E = e)P(H = v_3 \,|\, E = e)P(Wi = v_4 \,|\, E = e) \\ P(Wa = v_5 \,|\, E = e)P(F = v_6 \,|\, E = e)P(E = e)$$

E=0	E=1
p_0	p_1

P(S|E)	S=Sunny	S=Cloudy	S=Rainy
E=0	p_2	p_3	p_4
E=1	p_5	p_6	p_7

P(A|E)	A=Warm	A=Cold
E=0	p_8	p_9
E=1	p_{10}	p_{11}

P(H|E)	H=Normal	H=High
E=0	p_{12}	p_{13}
E=1	p_{14}	p_{15}

P(Wi|E)	Wi=Strong	Wi=Weak
E=0	p_{16}	p_{17}
E=1	p_{18}	p_{19}

P(Wa|E)	Wa=Warm	Wa=Cool
E=0	p_{20}	p_{21}
E=1	p_{22}	p_{23}

P(F|E)	F=Same	F=Change
E=0	p_{24}	p_{25}
E=1	p_{26}	p_{27}

E.g., <Rainy,Cold,High,Weak,Warm,Same>

$$\arg\max_{e\in\{0,1\}} P(S = Rainy \mid E = e)P(A = Cold \mid E = e)P(H = High \mid E = e)P(Wi = Weak \mid E = e)$$
$$P(Wa = Warm \mid E = e)P(F = Same \mid E = e)P(E = e)$$

for e=0 $\quad p_4 \cdot p_9 \cdot p_{13} \cdot p_{17} \cdot p_{20} \cdot p_{24} \cdot p_0$

return the label for which the product is larger

for e=1 $\quad p_7 \cdot p_{11} \cdot p_{15} \cdot p_{19} \cdot p_{22} \cdot p_{26} \cdot p_1$

This is called the **Naive Bayes** (NB) classifier

E=0	E=1
p_0	p_1

P(S|E)	S=Sunny	S=Cloudy	S=Rainy
E=0	p_2	p_3	p_4
E=1	p_5	p_6	p_7

P(A|E)	A=Warm	A=Cold
E=0	p_8	p_9
E=1	p_{10}	p_{11}

P(H|E)	H=Normal	H=High
E=0	p_{12}	p_{13}
E=1	p_{14}	p_{15}

P(Wi|E)	Wi=Strong	Wi=Weak
E=0	p_{16}	p_{17}
E=1	p_{18}	p_{19}

P(Wa|E)	Wa=Warm	Wa=Cool
E=0	p_{20}	p_{21}
E=1	p_{22}	p_{23}

P(F|E)	F=Same	F=Change
E=0	p_{24}	p_{25}
E=1	p_{26}	p_{27}

Given a training sample S of size m, we can estimate the ML parameters by counting:
let c(X=x) be the number of examples in S where X=x

Class prior = relative frequency of labels w.r.t. the full data, i.e.,

$$P(E=0) = \frac{c(E=0)}{m} \text{ and } P(E=1) = \frac{c(E=1)}{m}$$

Conditional probabilities = relative frequencies w.r.t. the examples of the given class, e.g.,

$$P(S = Sunny \,|\, E=0) = \frac{c(S = Sunny \wedge E = 0)}{c(E=0)}$$

E=0	E=1
p_0	p_1

P(S|E)	S=Sunny	S=Cloudy	S=Rainy
E=0	p_2	p_3	p_4
E=1	p_5	p_6	p_7

P(A|E)	A=Warm	A=Cold
E=0	p_8	p_9
E=1	p_{10}	p_{11}

P(H|E)	H=Normal	H=High
E=0	p_{12}	p_{13}
E=1	p_{14}	p_{15}

P(Wi|E)	Wi=Strong	Wi=Weak
E=0	p_{16}	p_{17}
E=1	p_{18}	p_{19}

P(Wa|E)	Wa=Warm	Wa=Cool
E=0	p_{20}	p_{21}
E=1	p_{22}	p_{23}

P(F|E)	F=Same	F=Change
E=0	p_{24}	p_{25}
E=1	p_{26}	p_{27}

Sky	AirTemp	Humidity	Wind	Water	Forecast	Enjoy
Sunny	Warm	Normal	Weak	Cool	Change	0
Sunny	Cold	High	Weak	Cool	Change	0
Rainy	Warm	Normal	Strong	Warm	Change	0
Cloudy	Warm	High	Strong	Warm	Same	1
Rainy	Warm	High	Weak	Cool	Same	1
Rainy	Warm	Normal	Weak	Warm	Change	0
Rainy	Cold	Normal	Weak	Cool	Change	1
Cloudy	Cold	High	Weak	Warm	Change	1
Sunny	Warm	High	Weak	Warm	Change	1
Sunny	Cold	Normal	Strong	Warm	Same	1
Cloudy	Warm	Normal	Strong	Cool	Change	0
Sunny	Cold	High	Strong	Cool	Same	0
Rainy	Warm	Normal	Weak	Cool	Change	1
Rainy	Warm	High	Strong	Cool	Change	0
Rainy	Cold	Normal	Strong	Warm	Change	0
Rainy	Warm	Normal	Weak	Warm	Same	1
Cloudy	Cold	Normal	Strong	Cool	Change	0
Cloudy	Cold	High	Strong	Cool	Change	0
Sunny	Cold	Normal	Strong	Warm	Same	1
Sunny	Warm	High	Weak	Cool	Change	1

E=0	E=1

P(S|E)	S=Sunny	S=Cloudy	S=Rainy
E=0			
E=1			

P(A|E)	A=Warm	A=Cold
E=0		
E=1		

and analogously for the remaining tables …

Naive Bayes

- Robust to irrelevant attributes
- Robust to isolated noisy data points
- Fewer parameters than full joint distribution, requiring less training data
- Often good classification results in practice, even if conditional independence assumption not justified

Homework

- Revise and practice the material seen today; it is the foundation for the rest of the module.
- Read chapter 1 of Barber's book and work through the examples (for the discrete case) it provides.
- Further exercises to help with this will be available on Learning Central after the lecture.

Reading Material

Note: the books all use slightly different notation to talk about the same concepts

- Today:
  - Understanding Machine Learning: parts of chapter 24
  - Russell & Norvig: chapter 13; parts of chapter 20
  - Barber: chapter 1; parts of chapter 10
- Next week:
  - Russell & Norvig: 14.1 & 14.2
  - Barber: chapters 2 & 3