THE UNIVERSITY of EDINBURGH
informatics

Applied Machine Learning (AML)

Representing Data

Oisin Mac Aodha • Siddharth N.

Representing Data

Core Questions

- What task am I trying to solve?
- How should I model the problem?
- How should I represent my data?
- How can I estimate the parameters of my model?
- How should I measure the performance of my model?

Core Questions

- What task am I trying to solve?
- How should I model the problem?
- How should I represent my data?
- How can I estimate the parameters of my model?
- How should I measure the performance of my model?

Representing Data

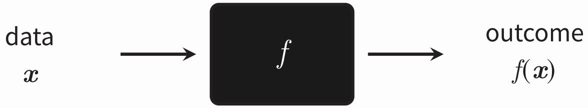
$$\begin{array}{c} \text{data} \\ \boldsymbol{x} \end{array} \longrightarrow \boxed{f} \longrightarrow \begin{array}{c} \text{outcome} \\ f(\boldsymbol{x}) \end{array}$$

Representing Data

data $\boldsymbol{x}$ $\longrightarrow$ $\boxed{f}$ $\longrightarrow$ outcome $f(\boldsymbol{x})$

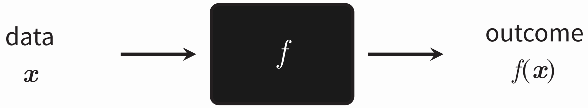
- How do we represent data ($\boldsymbol{x}$) mathematically?

Representing Data



- How do we represent data ($\boldsymbol{x}$) mathematically?
- What is the data, and what outcome we want?

Representing Data



- How do we represent data ($\boldsymbol{x}$) mathematically?
- What is the data, and what outcome we want?
    - If $\boldsymbol{x}$ is a person, are they eligible for a loan?

Representing Data

$$\underset{\boldsymbol{x}}{\text{data}} \longrightarrow \boxed{f} \longrightarrow \underset{f(\boldsymbol{x})}{\text{outcome}}$$

- How do we represent data ($\boldsymbol{x}$) mathematically?
- What is the data, and what outcome we want?
  - If $\boldsymbol{x}$ is a person, are they eligible for a loan?
  - If $\boldsymbol{x}$ is a chest scan, does the person have a tumour?

Feature-Value Pairs

Generic way to formulate representations

Feature-Value Pairs

Generic way to formulate representations

Characteristics

- what values do they take?
- what features/attributes to pick?

Feature-Value Pairs

Generic way to formulate representations

Characteristics

- what values do they take?
- what features/attributes to pick?

Categorical: {'red', 'blue', ...}

Ordinal: {'dislike', 'neutral', 'like'}

Numeric: $-3.14, 0.2, 1.4, \ldots$

Feature-Value Pairs

Generic way to formulate representations

Characteristics

- what values do they take?
- what features/attributes to pick?

Data: scale, similarity
structured vs. unstructured

Task: relevance, noise

Representing Data

What values can features take?

Categorical Features

- Each instance falls into one of a set of categories
  - E.g. musical *genre*: {‘classical’, ‘rock’, ‘jazz’, ‘techno’}
  - categories are mutually exclusive

Categorical Features

- Each instance falls into one of a set of categories
  - E.g. musical *genre*: {‘classical’, ‘rock’, ‘jazz’, ‘techno’}
  - categories are mutually exclusive
- Typically encoded as numbers that *index* into the set
  - E.g. ‘rock’ = 2
  - no natural ordering to the categories
  - no notion of ‘closeness’; only equality testing ($=, \neq$) is meaningful

Ordinal Features

- Each instance falls into one of a set of categories

Ordinal Features

- Each instance falls into one of a set of categories

- There is a *natural ordering* to the categories
  - E.g. marking scale: {‘poor’, ‘fair’, ‘good’, ‘excellent’}
  - categories are increasing (or decreasing) in some space

Ordinal Features

- Each instance falls into one of a set of categories

- There is a *natural ordering* to the categories
  - E.g. marking scale: {'poor', 'fair', 'good', 'excellent'}
  - categories are increasing (or decreasing) in some space

- Typically encoded as numbers that *preserve* ordering
  - E.g. 'poor' = 1, …, 'excellent' = 4
  - meaningful to compare $(<, =, >)$ values
  - ***not*** meaningful for other operations (e.g. add, multiply, …)

Numeric Features

- Integers ($\mathbb{Z}, \mathbb{N}$) or real numbers ($\mathbb{R}$)
  - integers viewed as *implicitly quantizing* continuous values

Numeric Features

- Integers $(\mathbb{Z}, \mathbb{N})$ or real numbers $(\mathbb{R})$
  - integers viewed as *implicitly quantizing* continuous values

- Has the whole gamut of characteristics
  - E.g. height of people: {165cm, 170cm, 188cm, …}
  - comparison $(<, =, >)$, closeness $|3.14 - 3.00|$, functions (e.g. `mean`, `variance`).
  - usually bounded and normalised: e.g. zero mean, unit variance

Numeric Features

- Integers $(\mathbb{Z}, \mathbb{N})$ or real numbers $(\mathbb{R})$
  - integers viewed as *implicitly quantizing* continuous values

- Has the whole gamut of characteristics
  - E.g. height of people: {165cm, 170cm, 188cm, ...}
  - comparison $(<, =, >)$, closeness $|3.14 - 3.00|$, functions (e.g. `mean`, `variance`).
  - usually bounded and normalised: e.g. zero mean, unit variance

- Can extend to higher order features
  - $\boldsymbol{x} \in \mathbb{R}^D$ for $D = 1$ is scalar, $D > 1$ is a vector
  - $\boldsymbol{x} \in \mathbb{R}^{D_1 \times \cdots \times D_N}$ for $N = 1$ is a vector, $N = 2$ is a matrix, ...

Examples of Representing Data

Example: Structured / Tabular Data

Should an applicant be given a loan?

Example: Structured / Tabular Data

Should an applicant be given a loan?

- Categorical
  - purpose: {‘car’, ‘home’, ‘education’, ‘business’}
  - personal: {‘single’, ‘married’, ‘divorced’, ‘separated’}

Example: Structured / Tabular Data

Should an applicant be given a loan?

- Categorical
  - purpose: {‘car’, ‘home’, ‘education’, ‘business’}
  - personal: {‘single’, ‘married’, ‘divorced’, ‘separated’}
- Ordinal
  - savings per month: {0, <100, 100—500, 500—1000, >1000}
  - current employment period: {‘unemployed’, <1yr, 1—4yrs, >4yrs}

Example: Structured / Tabular Data

Should an applicant be given a loan?

- Categorical
  - purpose: {‘car’, ‘home’, ‘education’, ‘business’}
  - personal: {‘single’, ‘married’, ‘divorced’, ‘separated’}
- Ordinal
  - savings per month: {0, <100, 100—500, 500—1000, >1000}
  - current employment period: {‘unemployed’, <1yr, 1—4yrs, >4yrs}
- Numeric
  - loan amount: e.g. £1000
  - loan interest: e.g. 5%

Example: Structured / Tabular Data

Should an applicant be given a loan?

- Categorical
  - purpose: {‘car’, ‘home’, ‘education’, ‘business’}
  - personal: {‘single’, ‘married’, ‘divorced’, ‘separated’}
- Ordinal
  - savings per month: {0, <100, 100—500, 500—1000, >1000}
  - current employment period: {‘unemployed’, <1yr, 1—4yrs, >4yrs}
- Numeric
  - loan amount: e.g. £1000
  - loan interest: e.g. 5%

Each applicant can have different values for these features.

Example: Text Data

Is this email spam or not?

Example: Text Data

Is this email spam or not?

- Represent email text as feature *vector* $\boldsymbol{x} = [x_1, \ldots, x_D]$

Example: Text Data

Is this email spam or not?

- Represent email text as feature *vector* $\boldsymbol{x} = [x_1, \ldots, x_D]$
- Use binary *categorical* features $x_d \in \{0, 1\}$ to indicate presence of a word

Example: Text Data

Is this email spam or not?

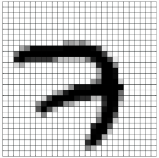
- Represent email text as feature *vector* $\boldsymbol{x} = [x_1, \ldots, x_D]$
- Use binary *categorical* features $x_d \in \{0, 1\}$ to indicate presence of a word
- Given the following vocabulary we can represent data as:

{ ‘password’, ‘review’, ‘send’, ‘us’, ‘your’, ‘account’ }

id	email	feature	status
1	“send us your password”	$[1, 0, 1, 1, 1, 0]$	spam
2	“send us review”	$[0, 1, 1, 1, 0, 0]$	ham
3	“review your account”	$[0, 1, 0, 0, 1, 1]$	ham
4	“review us”	$[0, 1, 0, 1, 0, 0]$	spam
5	“send your password”	$[1, 0, 1, 0, 1, 0]$	spam
6	“send us your account”	$[0, 0, 1, 1, 1, 1]$	spam

Example: Image Data

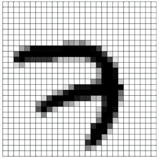
Pixels



- each pixel as separate feature
- numeric: degree of pixel “blackness”
- ordinal: binary $\in \{0, 1\}$

Example: Image Data

Pixels



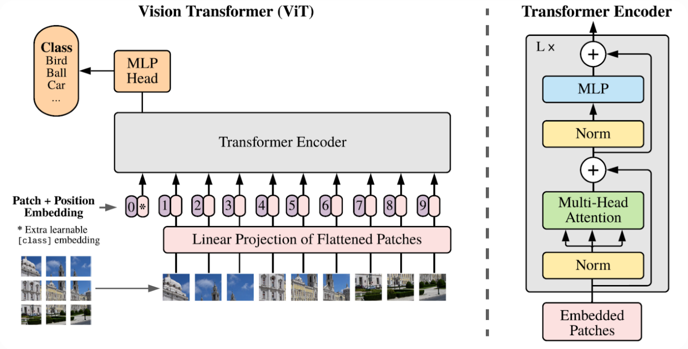
- each pixel as separate feature
- numeric: degree of pixel “blackness”
- ordinal: binary $\in \{0, 1\}$

Regions

- regions as separate features
- categorical: majority pixel class
- numeric: average pixel colour

Modern Representations

Choose a basic set of attributes, say image "patches"



[Dosovitskiy et al, An Image is Worth 16×16 Words, ICLR 2021]

Learn what values for features helps with doing the **task** well.

[R. A. Brooks, Intelligence without representation, *Artificial Intelligence* **47(1)**, 1991]

Representing Data: Summary

Representing Data: Summary

- Consider both data and task

data: what kinds of features to use

task: subset of features,
type of values

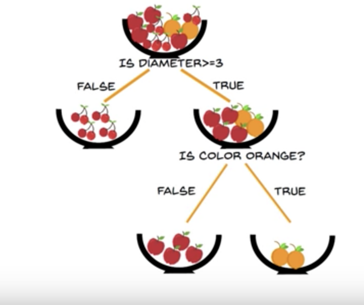
Representing Data: Summary

- Consider both data and task
- Consider what kind of model you want

Representing Data: Summary

- Consider both data and task
- Consider what kind of model you want
  - nuanced, interpretable

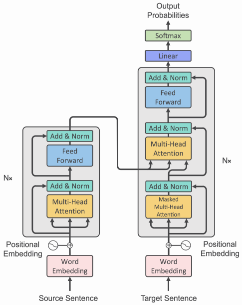
Constrained decision making over meaningful features



Representing Data: Summary

- Consider both data and task
- Consider what kind of model you want
  - nuanced, interpretable
  - scalable, performant

Opaque decision making over *learnt* task-specific features

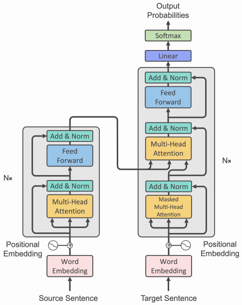


Representing Data: Summary

- Consider both data and task
- Consider what kind of model you want
  - nuanced, interpretable
  - scalable, performant

Powerful machine-learning models can be a *big* hammer

Opaque decision making over *learnt* task-specific features



Representing Data: Summary

- Consider both data and task
- Consider what kind of model you want
  - nuanced, interpretable
  - scalable, performant

Powerful machine-learning models can be a *big* hammer …is your problem a nail?

Opaque decision making over *learnt* task-specific features