ORIE 4741: Learning with Big Messy Data

Feature Engineering

Professor Udell
Operations Research and Information Engineering
Cornell

September 16, 2021
Announcements 9/16/21

- section posted
- bonus section from last year: linear algebra review
- hw1 due today at 9:15am
- form project groups by this Sunday. see https://people.orie.cornell.edu/mru8/orie4741/projects.html
- looking for a project group? post your idea on zulip in the #project channel
Poll

I want more office hours

A. Thursday
B. Friday
C. Saturday
D. Sunday
E. Monday – Wednesday
Outline

Feature engineering

Polynomial transformations

Boolean, nominal, ordinal, text, ...

Time series
To fit a linear model (linear in parameters $w$)

- pick a transformation $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$
- predict $y$ using a linear function of $\phi(x)$

$$h(x) = w^T \phi(x) = \sum_{i=1}^{d} w_i (\phi(x))_i$$

- we want $h(x_i) \approx y_i$ for every $i = 1, \ldots, n$
Linear models

To fit a linear model (= linear in parameters $w$)

- pick a transformation $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$
- predict $y$ using a linear function of $\phi(x)$

$$
    h(x) = w^T \phi(x) = \sum_{i=1}^{d} w_i(\phi(x))_i
$$

- we want $h(x_i) \approx y_i$ for every $i = 1, \ldots, n$

**Q:** why do we want a model linear in the parameters $w$?
Linear models

To fit a linear model (= linear in parameters $w$)

- pick a transformation $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$
- predict $y$ using a linear function of $\phi(x)$

$$h(x) = w^T \phi(x) = \sum_{i=1}^{d} w_i (\phi(x))_i$$

- we want $h(x_i) \approx y_i$ for every $i = 1, \ldots, n$

**Q:** why do we want a model linear in the parameters $w$?

**A:** because the optimization problems are easy to solve!

* e.g., just use least squares.
Feature engineering

How to pick $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$?

- so response $y$ will depend linearly on $\phi(x)$
- so $d$ is not too big
Feature engineering

How to pick $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$?

- so response $y$ will depend linearly on $\phi(x)$
- so $d$ is not too big

if you think this looks like a hack: you’re right
Feature engineering

examples:

- adding offset
- standardizing features
- polynomial fits
- products of features
- autoregressive models
- local linear regression
- transforming Booleans
- transforming ordinals
- transforming nominals
- transforming images
- transforming text
- concatenating data
- all of the above

https://xkcd.com/2048/
Outline

Feature engineering

Polynomial transformations

Boolean, nominal, ordinal, text, ...

Time series
Adding offset

\[ X = \mathbb{R}^{d-1} \]

let \( \phi(x) = (x, 1) \)

now \( h(x) = w^T \phi(x) = w_{1:d-1}^T x + w_d \)
Fitting a polynomial

\[ x \in \mathbb{R} \]

let

\[ \phi(x) = (1, x, x^2, x^3, \ldots, x^{d-1}) \]

be the vector of all monomials in \( x \) of degree < \( d \)

now

\[ h(x) = w^T \phi(x) = w_1 + w_2x + w_3x^2 + \cdots + w_d x^{d-1} \]
Demo: Linear models

https://github.com/ORIE4741/demos
IMHE and the cubic fit

The ‘cubic fit’ can depend on the data you use

[Graph showing cubic fit and time frames]

Source: IHME, Johns Hopkins University, Post analysis

https://www.washingtonpost.com/politics/2020/05/05/white-houses-self-serving-approach-estimating-deadliness-
Fitting a multivariate polynomial

- $X = \mathbb{R}^2$
- pick a maximum degree $k$
- let

$$\phi(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3 \ldots, x_2^k)$$

be the vector of all monomials in $x_1$ and $x_2$ of degree $\leq k$
- now $h(x) = w^T \phi(x)$ can fit any polynomial of degree $\leq k$ in $X$
Fitting a multivariate polynomial

- $\mathcal{X} = \mathbb{R}^2$
- pick a maximum degree $k$
- let
  
  $$\phi(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3 \ldots, x_2^k)$$

be the vector of all monomials in $x_1$ and $x_2$ of degree $\leq k$

- now $h(x) = w^T \phi(x)$ can fit any polynomial of degree $\leq k$
  in $\mathcal{X}$

and similarly for $\mathcal{X} = \mathbb{R}^d \ldots$
Polynomial classification

![Polynomial classification graph]
Example 1: multivariate polynomial classification

- $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{-1, 1\}$
- let $\phi(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)$ be the vector of all monomials of degree $\leq 2$
- now let $h(x) = \text{sign}(w^T \phi(x))$

**Q:** if $h(x) = \text{sign}(-30 - 9x_1 + 2x_2 + x_1^2 + x_2^2)$, what is $\{x : h(x) = 1\}$?

- A. a circle
- B. an ellipse
- C. a line
- D. a hyperbola
- E. a half-plane
Example 2: multivariate polynomial classification

- $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{-1, 1\}$
- let
  \[ \phi(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2) \]
  be the vector of all monomials of degree $\leq 2$
- now let $h(x) = \text{sign}(w^T \phi(x))$

**Q:** if $h(x) = \text{sign}(-5 - 3x_1 + 2x_2 + x_1^2 - x_1x_2 + 5x_2^2)$, what is $\{x : h(x) = 1\}$?

A. a circle
B. an ellipse
C. a line
D. a hyperbola
E. a half-plane
Example 3: multivariate polynomial classification

\( \mathcal{X} = \mathbb{R}^2, \mathcal{Y} = \{-1, 1\} \)

let

\[ \phi(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2) \]

be the vector of all monomials of degree \( \leq 2 \)

now let \( h(x) = \text{sign}(w^T \phi(x)) \)

Q: if \( h(x) = \text{sign}(-5 - 3x_1 + 2x_2 + x_1^2 - x_1x_2 + 5x_2^2) \), what is \( \{x : h(x) = 1\} \)?

A. a circle
B. an ellipse
C. a line
D. a hyperbola
E. a half-plane
Example 3: multivariate polynomial classification

- $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{-1, 1\}$
- let
  \[
  \phi(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)
  \]
  be the vector of all monomials of degree $\leq 2$
- now let $h(x) = \text{sign}(w^T\phi(x))$

Q: if $h(x) = \text{sign}(-5 - 3x_1 + 2x_2 + x_1^2 - x_1x_2 - 2x_2^2)$, what is $\{x : h(x) = 1\}$?

A. a circle
B. an ellipse
C. a line
D. a hyperbola
E. a half-plane
Outline

Feature engineering

Polynomial transformations

Boolean, nominal, ordinal, text, ...

Time series
Notation: boolean indicator function

define

\[ \mathbb{1}(\text{statement}) = \begin{cases} 1 & \text{statement is true} \\ 0 & \text{statement is false} \end{cases} \]

eamples:

- \( \mathbb{1}(1 < 0) = 0 \)
- \( \mathbb{1}(17 = 17) = 1 \)
Boolean variables

- $\mathcal{X} = \{\text{true, false}\}$
- let $\phi(x) = 1(x)$
Boolean expressions

- $\mathcal{X} = \{\text{true, false}\}^2 =\{(\text{true, true}), (\text{true, false}), (\text{false, true}), (\text{false, false})\}$.
- let $\phi(x) = [1(x_1), 1(x_2), 1(x_1 \text{ and } x_2), 1(x_1 \text{ or } x_2)]$
- equivalent: polynomials in $[1(x_1), 1(x_2)]$ span the same space
- encodes logical expressions!
Nominal values: one-hot encoding

- nominal data: e.g., $\mathcal{X} = \{\text{apple, orange, banana}\}$
- let

$$\phi(x) = [\mathbb{1}(x = \text{apple}), \mathbb{1}(x = \text{orange}), \mathbb{1}(x = \text{banana})]$$

- called **one-hot encoding**: only one element is non-zero
Nominal values: one-hot encoding

- Nominal data: e.g., $\mathcal{X} = \{\text{apple, orange, banana}\}$
- Let
  
  $\phi(x) = [\mathbb{1}(x = \text{apple}), \mathbb{1}(x = \text{orange}), \mathbb{1}(x = \text{banana})]$

- Called **one-hot encoding**: only one element is non-zero

**extension**: sets
Nominal values: look up features!

why not use other information known about each item?

- $\mathcal{X} = \{\text{apple, orange, banana}\}$
  - price, calories, weight, ...
- $\mathcal{X} = \text{zip code}$
  - average income, temperature in July, walk score, % residential, ...
- ...

database lingo: **join** tables on nominal value
Ordinal values: real encoding

- ordinal data: \( e.g., \)
  \[ \mathcal{X} = \{ \text{Stage I, Stage II, Stage III, Stage IV} \} \]
- let
  \[ \phi(x) = \begin{cases} 
  1, & x = \text{Stage I} \\
  2, & x = \text{Stage II} \\
  3, & x = \text{Stage III} \\
  4, & x = \text{Stage IV} 
\end{cases} \]
- default encoding
Ordinal values: real encoding

- $\mathcal{X} = \{\text{Stage I, Stage II, Stage III, Stage IV}\}$
- $\mathcal{Y} = \mathbb{R}$, number of years lived after diagnosis
- use real encoding $\phi$ to transform ordinal data
- fit linear model with offset to predict $y$ as $w\phi(x) + b$

Suppose model predicts a person diagnosed with Stage II cancer will survive 2 more years, and a person diagnosed with Stage I cancer will survive 4 more years.
Ordinal values: real encoding

- $\mathcal{X} = \{\text{Stage I, Stage II, Stage III, Stage IV}\}$
- $\mathcal{Y} = \mathbb{R}$, number of years lived after diagnosis
- use real encoding $\phi$ to transform ordinal data
- fit linear model with offset to predict $y$ as $w\phi(x) + b$

Suppose model predicts a person diagnosed with Stage II cancer will survive 2 more years, and a person diagnosed with Stage I cancer will survive 4 more years.

**Q:** What is $w$? $b$?
Ordinal values: real encoding

- $\mathcal{X} = \{\text{Stage I, Stage II, Stage III, Stage IV}\}$
- $\mathcal{Y} = \mathbb{R}$, number of years lived after diagnosis
- use real encoding $\phi$ to transform ordinal data
- fit linear model with offset to predict $y$ as $w\phi(x) + b$

Suppose model predicts a person diagnosed with Stage II cancer will survive 2 more years, and a person diagnosed with Stage I cancer will survive 4 more years.

**Q:** What is $w$? $b$?

A. $b = 6$, $w = -2$
B. $b = 2$, $w = 0$
C. $b = 6$, $w = 2$
D. $b = 0$, $w = -2$
Ordinal values: real encoding

- $\mathcal{X} = \{\text{Stage I, Stage II, Stage III, Stage IV}\}$
- $\mathcal{Y} = \mathbb{R}$, number of years lived after diagnosis
- use real encoding $\phi$ to transform ordinal data
- fit linear model with offset to predict $y$ as $w\phi(x) + b$

Suppose model predicts a person diagnosed with Stage II cancer will survive 2 more years, and a person diagnosed with Stage I cancer will survive 4 more years: $b = 6$, $w = -2$. 

Q: How long does the model predict a person with Stage IV cancer will survive?

A. 6 years
B. 2 years
C. 0 years
D. -2 years
**Ordinal values: real encoding**

- \( \mathcal{X} = \{ \text{Stage I, Stage II, Stage III, Stage IV} \} \)
- \( \mathcal{Y} = \mathbb{R} \), number of years lived after diagnosis
- Use real encoding \( \phi \) to transform ordinal data
- Fit linear model with offset to predict \( y \) as \( w\phi(x) + b \)

Suppose model predicts a person diagnosed with Stage II cancer will survive 2 more years, and a person diagnosed with Stage I cancer will survive 4 more years: \( b = 6 \), \( w = -2 \).

**Q:** How long does the model predict a person with Stage IV cancer will survive?
Ordinal values: real encoding

- $\mathcal{X} = \{\text{Stage I, Stage II, Stage III, Stage IV}\}$
- $\mathcal{Y} = \mathbb{R}$, number of years lived after diagnosis
- use real encoding $\phi$ to transform ordinal data
- fit linear model with offset to predict $y$ as $w\phi(x) + b$

Suppose model predicts a person diagnosed with Stage II cancer will survive 2 more years, and a person diagnosed with Stage I cancer will survive 4 more years: $b = 6$, $w = -2$.

**Q:** How long does the model predict a person with Stage IV cancer will survive?

A. 6 years
B. 2 years
C. 0 years
D. -2 years
Ordinal values: boolean encoding

- ordinal data: e.g.,
  \( \mathcal{X} = \{ \text{Stage I, Stage II, Stage III, Stage IV} \} \)
- let

  \[
  \phi(x) = [\mathbb{1}(x \geq \text{Stage II}), \mathbb{1}(x \geq \text{Stage III}), \mathbb{1}(x \geq \text{Stage IV})]
  \]
Ordinal values: boolean encoding

- $\mathcal{X} = \{\text{Stage I}, \text{Stage II}, \text{Stage III}, \text{Stage IV}\}$
- $\mathcal{Y} = \mathbb{R}$, number of years lived after diagnosis
- define transformation $\phi : \mathcal{X} \rightarrow \mathbb{R}$ as

$$\phi(x) = [\mathbb{1}(x \geq \text{Stage II}), \mathbb{1}(x \geq \text{Stage III}), \mathbb{1}(x \geq \text{Stage IV})]$$

- fit linear model with offset to predict $y$ as $w^\top \phi(x) + b$

Suppose model predicts a person diagnosed with Stage II cancer will survive 2 more years, and a person diagnosed with Stage I cancer will survive 4 more years.
Ordinal values: boolean encoding

- $\mathcal{X} = \{\text{Stage I, Stage II, Stage III, Stage IV}\}$
- $\mathcal{Y} = \mathbb{R}$, number of years lived after diagnosis
- define transformation $\phi : \mathcal{X} \rightarrow \mathbb{R}$ as
  \[
  \phi(x) = [\mathbb{1}(x \geq \text{Stage II}), \mathbb{1}(x \geq \text{Stage III}), \mathbb{1}(x \geq \text{Stage IV})]
  \]
- fit linear model with offset to predict $y$ as $\mathbf{w}^\top \phi(x) + b$

Suppose model predicts a person diagnosed with Stage II cancer will survive 2 more years, and a person diagnosed with Stage I cancer will survive 4 more years.

**Q:** What is $w$? $b$?
Ordinal values: boolean encoding

- $\mathcal{X} = \{\text{Stage I, Stage II, Stage III, Stage IV}\}$
- $\mathcal{Y} = \mathbb{R}$, number of years lived after diagnosis
- define transformation $\phi : \mathcal{X} \rightarrow \mathbb{R}$ as

$$\phi(x) = [\mathbb{1}(x \geq \text{Stage II}), \mathbb{1}(x \geq \text{Stage III}), \mathbb{1}(x \geq \text{Stage IV})]$$

- fit linear model with offset to predict $y$ as $w^\top \phi(x) + b$

Suppose model predicts a person diagnosed with Stage II cancer will survive 2 more years, and a person diagnosed with Stage I cancer will survive 4 more years.

**Q:** What is $w$? $b$?

**A:** $b = 4$, $w_1 = -2$, $w_2$ and $w_3$ not determined
Ordinal values: boolean encoding

- $\mathcal{X} = \{\text{Stage I}, \text{Stage II}, \text{Stage III}, \text{Stage IV}\}$
- $\mathcal{Y} = \mathbb{R}$, number of years lived after diagnosis
- define transformation $\phi : \mathcal{X} \rightarrow \mathbb{R}$ as

  $$\phi(x) = [\mathbb{1}(x \geq \text{Stage II}), \mathbb{1}(x \geq \text{Stage III}), \mathbb{1}(x \geq \text{Stage IV})]$$

- fit linear model with offset to predict $y$ as $w^\top \phi(x) + b$

Suppose model predicts a person diagnosed with Stage II cancer will survive 2 more years, and a person diagnosed with Stage I cancer will survive 4 more years.

Q: What is $w$? $b$?
A: $b = 4$, $w_1 = -2$, $w_2$ and $w_3$ not determined

Q: How long does the model predict a person with Stage IV cancer will survive?
Ordinal values: boolean encoding

- $\mathcal{X} = \{\text{Stage I, Stage II, Stage III, Stage IV}\}$
- $\mathcal{Y} = \mathbb{R}$, number of years lived after diagnosis
- define transformation $\phi : \mathcal{X} \rightarrow \mathbb{R}$ as
  
  $$\phi(x) = [\mathbb{1}(x \geq \text{Stage II}), \mathbb{1}(x \geq \text{Stage III}), \mathbb{1}(x \geq \text{Stage IV})]$$

- fit linear model with offset to predict $y$ as $w^T \phi(x) + b$

Suppose model predicts a person diagnosed with Stage II cancer will survive 2 more years, and a person diagnosed with Stage I cancer will survive 4 more years.

**Q:** What is $w$? $b$?

**A:** $b = 4$, $w_1 = -2$, $w_2$ and $w_3$ not determined

**Q:** How long does the model predict a person with Stage IV cancer will survive?

**A:** can’t say without more information
$X$ = sentences, documents, tweets, . . .

- **Bag of words** model (one-hot encoding):
  - pick set of words $\{w_1, \ldots, w_d\}$
  - $\phi(x) = [\mathbb{1}(x \text{ contains } w_1), \ldots, \mathbb{1}(x \text{ contains } w_d)]$
  - ignores order of words in sentence
$\mathcal{X} =$ sentences, documents, tweets, …

- **bag of words** model (one-hot encoding):
  - pick set of words $\{w_1, \ldots, w_d\}$
  - $\phi(x) = [\mathbb{1}(x \text{ contains } w_1), \ldots, \mathbb{1}(x \text{ contains } w_d)]$
  - ignores order of words in sentence

- **pre-trained neural networks**:
  - sentiment analysis: https://medium.com/@b.terryjack/nlp-pre-trained-sentiment-analysis-1eb52a9d742c
  - Universal Sentence Encoder (USE) embedding: https://colab.research.google.com/github/tensorflow/hub/blob/master/examples/colab/semantic_similarity_with_tf_hub_universal_encoder.ipynb
  - lots of others: https://modelzoo.co/
Neural networks: whirlwind primer

\[ \text{NN}(x) = \sigma(W_1\sigma(W_2 \ldots \sigma(W_\ell x)))) \]

- \( \sigma \) is a nonlinearity applied elementwise to a vector, e.g.
  - ReLU: \( \sigma(x) = \max(x, 0) \)
  - sigmoid: \( \sigma(x) = \log(1 + \exp(x)) \)
- each \( W \) is a matrix
- trained on very large datasets, e.g., Wikipedia, YouTube

![Deep Neural Network](image)

Figure 12.2: Deep network architecture with multiple layers.
Why not use deep learning?

Common carbon footprint benchmarks

<table>
<thead>
<tr>
<th>Activity</th>
<th>Carbon Equivalent (lbs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roundtrip flight b/w NY and SF (1 passenger)</td>
<td>1,984</td>
</tr>
<tr>
<td>Human life (avg. 1 year)</td>
<td>11,023</td>
</tr>
<tr>
<td>American life (avg. 1 year)</td>
<td>36,156</td>
</tr>
<tr>
<td>US car including fuel (avg. 1 lifetime)</td>
<td>126,000</td>
</tr>
<tr>
<td>Transformer (213M parameters) w/ neural architecture search</td>
<td>626,155</td>
</tr>
</tbody>
</table>

Chart: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

towards a solution: https://arxiv.org/abs/1907.10597
Outline

Feature engineering

Polynomial transformations

Boolean, nominal, ordinal, text, …

Time series
Fitting a time series

- given a time series \( x_t \in \mathbb{R}, t = 1, \ldots, T \)
- want to predict the value at the next time \( T + 1 \)
- input: time index \( t \) and time series \( x_{1:t} \) up to time \( t \)
- let
  \[
  \phi(t, x) = (x_{t-1}, x_{t-2}, \ldots, x_{t-d})
  \]
  (called the “lagged outcomes”)
- now \( h(x) = w^T \phi(x) = w_1 x_{t-1} + w_2 x_{t-2} + \cdots + w_d x^{t-d} \)

also called an auto-regressive (AR) model
Fitting a local model: local neighbors

- $\mathcal{X} = \mathbb{R}^m$
- pick a set of points $\mu_i \in \mathbb{R}^m$, $i = 1, \ldots, d$, radius $\delta \in \mathbb{R}$
- let $\phi(x) = [\mathbb{1}(\|x - \mu_1\|^2 \leq \delta), \ldots, \mathbb{1}(\|x - \mu_d\|^2 \leq \delta)]$

Often, $d = n$ and $\mu_i = x_i$
Fitting a local model: 1 nearest neighbor

- \( \mathcal{X} = \mathbb{R}^m \)
- pick a set of points \( \mu_i \in \mathbb{R}^m, i = 1, \ldots, d \)
- let \( \delta = \min(\|x - \mu_1\|^2, \ldots, \|x - \mu_d\|^2) \)
- let \( \phi(x) = [\mathbb{1}(\|x - \mu_1\|^2 \leq \delta), \ldots, \mathbb{1}(\|x - \mu_d\|^2 \leq \delta)] \)

often, \( d = n \) and \( \mu_i = x_i \)
Fitting a local model: smoothing

- $\mathcal{X} = \mathbb{R}^m$
- pick a set of points $\mu_i \in \mathbb{R}^m$, $i = 1, \ldots, d$, parameter $\alpha \in \mathbb{R}$
- let

$$\phi(x) = (\exp (-\alpha \| x - \mu_1 \|^2), \ldots, \exp (-\alpha \| x - \mu_d \|^2))$$

often, $d = n$ and $\mu_i = x_i$
Crime demo:
preprocessing: https://juliabox.com/notebooks/demos/Crime.ipynb
predicting: https://juliabox.com/notebooks/demos/Predicting%20crime.ipynb

COVID demo:
predicting: https://juliabox.com/notebooks/demos/Predicting%20COVID.ipynb
Review

- linear models are linear in the parameters $w$
- can fit many different models by picking feature mapping $\phi : \mathcal{X} \to \mathbb{R}^d$
What makes a good project?

- Clear outcome to predict
- Linear regression should do something interesting
- A data science project; not an NLP or Vision project
- New, interesting model; not a Kaggle competition