Love in the time of machine learning

Carlos Cotrini

ccarlos@inf.ethz.ch

Advanced Machine Learning 2019

ETH Zürich

October 2, 2019

The formula for everlasting love

Figure: Robyn Dawes

- American psychologist.
- 27 couples. 23 "happy", 3 "unhappy" and 1 "inconclusive".
- He proposed a simple predictor for marital happiness:

The formula for everlasting love

frequency of intimacy - frequency of quarrels

The formula for everlasting love

The formula for everlasting love is linear! More formally:

A linear model is powerful enough to predict marriage success!

How about predicting the amount of years a marriage will last?

Predicting number of years a marriage will last

1. Define which features you think define the number of years a marriage will last (e.g., frequency of quarrels, frequency of intimacy, etc...).
2. Collect a sufficiently large and representative sample X of marriages.

Formalization

Assume given $X \in \mathbb{R}^{N \times D}$ and $y \in \mathbb{R}^N$ such that:

- $x_i = (x_{i1}, x_{i2}, \dots, x_{iD})$ denotes X's i-th row.
- x_i represents all D features of marriage $i \leq N$ (e.g. frequency of lovemaking, frequency of quarrels, age difference, salary difference, etc...).
- y_i is the number of years of marriage i.

Objective

Find factors $\beta_1, \ldots, \beta_D \in \mathbb{R}$ such that for any marriage $x_i = (x_{i1}, \ldots, x_{iD})$, the value $\beta_1 x_{i1} + \ldots + \beta_D x_{iD}$ is very close to y_i.
In other words, find $\beta = (\beta_1, \ldots, \beta_D) \in \mathbb{R}^D$ such that for any marriage x_i, the value $x_i^\top \beta$ is very close to y_i.
More precisely,

$$\underset{\beta \in \mathbb{R}^D}{\arg\min} \quad \sum_{i \leq N} \left(x_i^\top \beta - y_i \right)^2.$$

Advertising the formula...

The formula doesn't sell well if it's too complex. Moreover, it risks "memorizing" the dataset.
How can we search for a simple one?

$$\arg\min_{\beta \in \mathbb{R}^D} \lambda \,(\#\text{ of non-zero entries in } \beta) + \sum_{i \leq N} \left(x_i^\top \beta - y_i \right)^2 .$$

But we want something more mathematical...

$$\arg\min_{\beta \in \mathbb{R}^D} \lambda \left\| \beta \right\|_1 + \sum_{i \leq N} \left(x_i^\top \beta - y_i \right)^2 .$$

Tibishirani proved that when λ is sufficiently high, then a minimizer of this is *sparse*.

But we would like something that we can optimize using only standard calculus...

$$\arg\min_{\beta \in \mathbb{R}^D} \lambda \|\beta\|_2^2 + \sum_{i \leq N} \left(x_i^\top \beta - y_i \right)^2 .$$

It is not guaranteed that the minimizer is sparse, but this regularization term is a good balance between a sparse minimizer and a minimizer that is easy to compute.

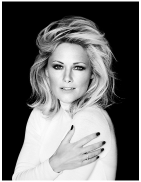
Commercials

If you are interested in a master thesis on ML for spring 2020,

`https://inf.ethz.ch/personal/ccarlos/projects.html`

Love personalities and clustering

(a) Helen Fisher



(b) Helene Fischer

Four love personalities

When you enter to chemistry.com you need to answer 60 questions like:

	Strongly disagree	Disagree	Agree	Strongly agree
I am always looking for new experiences.				
I think consistent routines keep life orderly and relaxing.				
I am more analytical and rational than most people.				
I am very sensitive to peoples feelings and needs.				

So for Helen Fisher, you are just a point in $\mathbb{R}^{60}$.

Four love personalities

When you enter to chemistry.com you need to answer 60 questions like:

	Strongly disagree	Disagree	Agree	Strongly agree
I am always looking for new experiences.				
I think consistent routines keep life orderly and relaxing.				
I am more analytical and rational than most people.				
I am very sensitive to peoples feelings and needs.				

She observed four clusters: explorers, builders, directors, and negotiators.

A model for love personalities

Assume given the dataset from chemistry.com as a matrix $X \in \mathbb{R}^{N \times D}$, where $x_i \in \mathbb{R}^D$ contains the answers for person $i \leq N$.
How can we model the sampling of x_i for person i?

1. Let $\mu_c \in \mathbb{R}^D$, with $c \leq C = 4$, be the expected answers given by an explorer, a builder, a director, and a negotiator.
2. Draw person i's personality type c with probability π_c. We model this personality type with a vector $M_i \in \{0,1\}^C$ such that $M_{ic} = 1$ and all other entries are zero.
3. Draw person i's vector of answers x_i by drawing from $\mathcal{N}(\mu_c, \Sigma_c)$, where Σ_c is a covariance matrix.

Objective

Discover, from the dataset X from chemistry.com, what are the values of μ_c, Σ_c, and π_c, for $c \leq C = 4$.

Maximum-likelihood approach

Let $\theta := \{\mu_1, \ldots, \mu_C, \Sigma_1, \ldots, \Sigma_C, \pi_1, \ldots, \pi_c\}$.
Our **objective** can be then formalized as

$$\arg\max_{\theta} \quad P(X \mid \theta).$$

For the sake of numerical stability, we optimize instead

$$\arg\max_{\theta} \quad \log P(X \mid \theta). \qquad \text{Incomplete-data log likelihood.}$$

For the sake of comparison, we will also consider

$$\arg\max_{\theta} \quad \log P(X, M \mid \theta). \qquad \text{Complete-data log likelihood.}$$

An important independence assumption

Any two different people in X are *independent*. Even when we know their personality types!

Does it make sense? What if there are twins? relatives?
We can't model everything...

Maximum-likelihood approach

We can show that

$$\log P(X \mid \theta) = \sum_{i \leq N} \log \left(\sum_{c \leq C} \pi_c \mathcal{N}\left(x_i \mid \mu_c, \Sigma_c\right) \right).$$

However, analytically maximizing this incomplete-data log likelihood is quite challenging.In contrast,

$$\log P(X, M \mid \theta) = \sum_{i \leq N} \sum_{c \leq C} M_{ic} \log \left(\pi_c \mathcal{N}\left(x_i \mid \mu_c, \Sigma_c\right)\right).$$

Analytical maximization of the complete-data log likelihood is much more manageable.

This happens often in mixtures from the exponential family.

The EM algorithm is a useful tool for maximizing incomplete-data log likelihoods. It leverages the fact that maximizing the complete-data log likelihood is relatively easy.

The EM algorithm

EM-algorithm

- Init: Initialize θ^o with random values.
- E-step: Compute $P\left(M \mid X, \theta^o\right)$.
- M-step: $\theta \leftarrow \arg\max_\theta \mathbb{E}_{P(M \mid X, \theta^o)}\left[\log P\left(X, M \mid \theta\right)\right]$.
- Repeat: If θ^o and θ are close enough, finish; otherwise, set $\theta^o \leftarrow \theta$ and go to [E-step].

This also works for mixtures of distributions from the exponential family.