Mathematical foundations.

Jonas Groschwitz, Antoine Venant

May 6, 2019

Reminder

The big picture

- Formalize problem as *learning* a function: $f : \mathbb{R}^n \mapsto \mathbb{R}^m$.
- Define a class of models. That, is a class of 'candidate' functions $g_\theta : \mathbb{R}^n \mapsto \mathbb{R}^m$ that we know how to compute.
  - $\theta \in \mathbb{R}^k$: parameters of the model.
- **Find** the model g_{θ^*} providing **the best approximation of** f **given available evidence.**

Question

- What is the best approximation (given available evidence)?
- (How do we find it?)

Maximum likelihood.

Classification.

- Observations: set of pairs $O = \{(x_1, y_{,1}) \dots (x_n, y_n)\}$.
- For each observed pair (x, y), $y \in \mathcal{Y} = \{c_1, \dots, c_k\}$, finite set of classes.
- Model: $p_\theta(y \mid x)$: for each possible input, determines conditional (predictive) probability over outcome in $\mathcal{Y}$.

Best model:

- Criterion for the quality of the model: how well does it account for observations: the higer $p_\theta(O)$, the better the model.
- Loss function (measure how 'bad' the model is): $\frac{1}{p_\theta(O)}$.

- Loss function (measure how 'bad' the model is): $\frac{1}{p_\theta(O)}$.

Potential issues:

1. What is actually $p_\theta(O)$? Model introduced above offers only conditional probability $p_\theta(y \mid x)$.
2. Theoretically, we should care about both unobserved and observed data. How exactly does this relate to maximizing the likelihood of data?
3. What if we want to use something other likelihood?

Potential issue 1

What is actually $p_\theta(O)$?

- Assume inputs follow some (paramter-independent) distribution.
- Mix that with the model predictive probability to obtain a parametrized distribution over observations.
- Here, for instance, assume inputs $x_1 \dots x_n$ are outcomes of some i.i.d. random variables with distribution p.

Likelihood of observations

In our example:

$$p_\theta(O) = \prod_{i=1}^{n} p(x_i) p_\theta(y_i \mid x_i)$$

Are we there yet?

Likelihood and loss

In our example:

$$p_\theta(O) = \prod_{i=1}^{n} p(x_i) p_\theta(y_i \mid x_i)$$

$$L(\theta) = \frac{1}{p_\theta(O)} = \prod_{i=1}^{n} \frac{1}{p(x_i) p_\theta(y_i \mid x_i)}$$

Best model's params: $argmin_\theta L(\theta)$

Wait a minute.

Are we there yet?

Likelihood and loss

In our example:

$$p_\theta(O) = \prod_{i=1}^{n} p(x_i) p_\theta(y_i \mid x_i)$$

$$L(\theta) = \frac{1}{p_\theta(O)} = \prod_{i=1}^{n} \frac{1}{p(x_i) p_\theta(y_i \mid x_i)}$$

Best model's params: $argmin_\theta L(\theta)$

Wait a minute.

- How do we compute L Without knowing p?

Are we there yet?

Likelihood and loss

In our example:

$$p_\theta(O) = \prod_{i=1}^{n} p(x_i)p_\theta(y_i \mid x_i)$$

$$L(\theta) = \frac{1}{p_\theta(O)} = \prod_{i=1}^{n} \frac{1}{p(x_i)p_\theta(y_i \mid x_i)}$$

Best model's params: $argmin_\theta L(\theta)$

Wait a minute.

- How do we compute L Without knowing p? We don't

Are we there yet?

Likelihood and loss

In our example:

$$p_\theta(O) = \prod_{i=1}^{n} p(x_i)p_\theta(y_i \mid x_i)$$

$$L(\theta) = \frac{1}{p_\theta(O)} = \prod_{i=1}^{n} \frac{1}{p(x_i)p_\theta(y_i \mid x_i)}$$

Best model's params: $argmin_\theta L(\theta)$

Wait a minute.

- How do we compute L Without knowing p? We don't
- Write $L(\theta) = \underbrace{\prod_{i=1}^{n} \frac{1}{p(x_i)}}_{\text{parameter-independent constant}} \times \underbrace{\prod_{i=1}^{n} \frac{1}{p_\theta(y_i|x_i)}}_{L'(\theta)}$.

Are we there yet?

Likelihood and loss

In our example:

$$p_\theta(O) = \prod_{i=1}^{n} p(x_i) p_\theta(y_i \mid x_i)$$

$$L(\theta) = \frac{1}{p_\theta(O)} = \prod_{i=1}^{n} \frac{1}{p(x_i) p_\theta(y_i \mid x_i)}$$

Best model's params: $argmin_\theta L(\theta)$

Wait a minute.

- How do we compute L Without knowing p? We don't
- Write $L(\theta) = \underbrace{\prod_{i=1}^{n} \frac{1}{p(x_i)}}_{\text{parameter-independent constant}} \times \underbrace{\prod_{i=1}^{n} \frac{1}{p_\theta(y_i|x_i)}}_{L'(\theta)}$.
- Minimizing L and L' is the same.

Potential issue 2

Theoretically, should'nt the 'best' model depend on unobserved data too?

- Of course we cannot *use* unavailable data in our search.
- But does not mean that the mathematical definition should not consider unobserved data.
- Assume data distributed following 'ground truth' distribution: $\hat{p}(x, y)$.
- Best model: model yielding the joint distribution 'closest' to $\hat{p}(x, y)$ *i.e.* $L(\theta) = DKL(\hat{p}||p_\theta)$.

Kullback-Leibler divergence

$$DKL(p||q) = \underbrace{\sum_{o \text{ possible data}} p(o) \times -log(q(o))}_{\text{Cross-entropy} H(p,q)} - \underbrace{\sum_o p(o) \times -log(p(o))}_{\text{entropy } H(p)}$$

Back to observations

Kullback-Leibler divergence

$$DKL(p||q) = \underbrace{\sum_{o \text{ possible data}} p(o) \times -log(q(o))}_{\text{Cross-entropy} H(p,q)} - \underbrace{\sum_{o} p(o) \times -log(p(o))}_{\text{entropy } H(p)}$$

- $H(p)$ does not depend q, so we can just search for q minimizing cross entropy.
- Back to practical consideration: must use available observations to approximate $H(p,q)$.
- Remark that $H(p,q) = \mathbb{E}_{X \sim p}[-log(q(X))]$ is an expectation (under p). Under some assumptions (*e.g.* i.i.d. observations, but not only) we can approximate

$$H(p,q) \sim \frac{1}{n} \sum_{i=1}^{n} -log(q(o_i))$$

using n observations $o_1, \ldots, o_n$.

Back to likelihood:

With notations from before:

- Ground truth: $\hat{p}(x, y)$.
- Model: $p_\theta(x, y) = \hat{p}(x)p_{\theta(y|x)}$.

$$H(\hat{p}, p_\theta) \sim \underbrace{\frac{1}{n}}_{\text{independent of } \theta} \underbrace{\sum_{i=1}^{n} -log(\hat{p}(x_i)p_\theta(y \mid x))}_{log\left(\prod_{i=1}^{n} \frac{1}{p(x_i)}\right)}$$

Last issue

What if we don't want to use likelihood (or DKL)?

- In the *DKL* setting 'badness' of the model measured by: $\mathbb{E}_{X\sim p}[-log(q(X))]$.
- Intuition: $-log(q(x)) = log(1/q(x))$: measures (badness of) performance of the model over one particular data point. Average over every data-point.
- The general case: consider arbitrary loss function $L(\theta, x)$ measuring performance over one data-point x.
- Minimize $\mathbb{E}_{X\sim p}[L(\theta, X)]$ for θ.

Finding the best model. First step.

- General method: (stochastic) gradient descent.
- Today, first part: gradients.
- Necessary condition for being the minimum of a differentiable function f: f has 0 derivative.

Derivatives: reminder (whiteboard)

Whiteboard

Critical points and local minima

Whiteboard

Special case: convex optimisation

Whiteboard

Common derived functions:

- $\forall n \in \mathbb{Z}$, $[x^n]' = nx^{n-1}$.
- $[log(x)]' = \frac{1}{x}$.
- $[e^x]' = e^x$.
- $[cos(x)]' = -sin(x)$
- $[sin(x)]' = cos(x)$

Reminder: property of derivatives

Chain rule:

For $f, g : \mathbb{R} \mapsto \mathbb{R}, A, B \subseteq \mathbb{R}$ s.t. g differentiable over A, f differentiable over $g(A)$,

$$[f \circ g]' = f' \circ g \times g'$$

Composition rules:

If f and g differentiable then

- $f + g$ differentiable and $(f + g)' = f' + g'$.
- f and g differentiable and $[f \times g]' = f'g + fg'$.

Partial derivatives

Say we have a function $f : \mathbb{R}^n \to \mathbb{R}$, e.g.

$$f(x, y) = x^2 y$$

Partial derivatives derive with respect to one input dimension, and fix all other inputs:

$$\frac{\partial}{\partial x} f(x, y) = 2xy$$
$$\frac{\partial}{\partial y} f(x, y) = x^2$$

Partial derivatives

The *gradient* is the $1 \times n$-dimensional vector of partial derivatives:

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{pmatrix}$$

Example: if again $f(x, y) = x^2 y$, then:

$$\nabla f(x, y) = \begin{pmatrix} 2xy & x^2 \end{pmatrix}$$

Deriving multi-dimensional functions

Say we have a function $f : \mathbb{R} \to \mathbb{R}^n$, e.g.

$$f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \end{pmatrix} = \begin{pmatrix} 5 - x \\ 3x^2 \end{pmatrix}$$

(We write $f_1(x) = 5 - 1$ and $f_2(x) = 3x^2$ for each dimension.)
Derivatives for multi-dimensional functions are just done separately for each dimension, and written in a Matrix called the *Jacobian*:

$$J_f(x) = \begin{pmatrix} f_1'(x) \\ f_2'(x) \end{pmatrix} = \begin{pmatrix} -1 \\ 6x \end{pmatrix}$$

The general case

Now for a function $f : \mathbb{R}^n \to \mathbb{R}^m$, we have the following Jacobian matrix:

$$J_f(x_1, \dots, x_n) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \dots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

Example

If the function is:

$$f(x,y) = \begin{pmatrix} 5 - x + 4y \\ x^2y^7 \end{pmatrix}$$

Then the Jacobian is:

$$J_f(x,y) = \begin{pmatrix} \frac{\partial f_1}{\partial x} & \frac{\partial f_1}{\partial y} \\ \frac{\partial f_2}{\partial x} & \frac{\partial f_2}{\partial y} \end{pmatrix} = \begin{pmatrix} -1 & 4 \\ 2xy^7 & 7x^2y^6 \end{pmatrix}$$

Generalized chain rule

Functions $f : \mathbb{R}^m \to \mathbb{R}^k$ and $g : \mathbb{R}^n \to \mathbb{R}^m$, want to derive their concatenation:

$$(f \circ g)(x_1, \ldots, x_n) = f(g(x_1, \ldots, x_n))$$

Then the Jacobian of the composed function is:

$$J_{f \circ g}(\mathbf{x}) = J_f(g(\mathbf{x})) \cdot J_g(\mathbf{x})$$