Statistics and Supervised Machine Learning: bridging the ‘gap’

Kamran Javid

Arabesque

kamran.javid@arabesque.com

January 4, 2023

Why is the link important?

- Supervised machine learning is crucial to the AI Engine
- In isolation, appears somewhat *ad-hoc*
- Today we will show it isn't!

Agenda

1. Preliminaries
   - Supervised machine learning
     - Theory
     - Example
   - Bayesian inference
     - Theory
     - Example
2. Putting the two together
   - Maximum posterior estimation
     - Theory
     - Example

Supervised machine learning

Definition (Supervised machine learning)

Given a *feature* matrix X, and *output* matrix Y, we want to learn a *mapping* f from X to Y Furthermore we can assert that our mapping is governed by some *model* $\mathcal{M}$, and depends explicitly on *model parameters* θ:

$$f : X, \theta \to Y | \mathcal{M} \tag{1}$$

Definition (Objective function)

Given a model $f(\cdot; \theta)$ and training data (X, Y), one can form an **objective function** $\mathcal{O}$ consisting of a **loss function** $\mathcal{L}$, and optionally a **weight decay** function $\mathcal{R}$:

$$\mathcal{O}(X, Y, \theta) = \tau_L \mathcal{L}(f(X; \theta), Y) + \tau_R \mathcal{R}(\theta), \tag{2}$$

Supervised machine learning continued

where τ_L and τ_R dictate the relative weighting of $\mathcal{L}$ and $\mathcal{R}$. The mapping is 'learned' by **minimising** $\mathcal{O}$ with respect to the model parameters θ:

$$\min_{\theta} \tau_L \mathcal{L}(f(X;\theta), Y) + \tau_R \mathcal{R}(\theta) \tag{3}$$

Roughly speaking, $\mathcal{L}$ measures the *similarity* between the true target Y, and the model's approximation, $f(X;\theta)$.
Weight decay serves to balance this in the optimisation, by 'penalising' models which are deemed to be overly *complex*

Supervised machine learning example

Illustrative example, predicting house prices:

- We have ten houses for which we know the *true price* (Y)
- We want to learn a mapping from *two features* (*land area* and *average temperature*, X) to true price
- We choose $\mathcal{M}$ such that our model of Y is a linear regression model, so $f(X;\theta) = X\theta$
- Where $X \in \mathbb{1}^{10\times 1} \times {\mathbb{R}^{+}}^{10\times 2}$, $Y, f(X;\theta) \in {\mathbb{R}^{+}}^{10\times 1}$, $\theta \in \mathbb{R}^{3\times 1}$
- Use *squared error* loss function $\mathcal{L} = (Y - f(X;\theta))^\top (Y - f(X;\theta))$
- We also use an *l2-norm squared* weight decay $\mathcal{R}(\theta) = \theta^\top \theta$

Definition (Bayesian inference)

Taking our definitions of $X, Y, f, \mathcal{M}$ and θ used for (1).
Say that we want to infer the model parameters θ conditional on the data (X, Y) and on our model assumptions. Unfortunately, nature has screwed us on this *inverse* problem. Thankfully, reverend *Bayes* saved the day:

$$\mathcal{P}(\theta|X, Y, f, \mathcal{M}) = \frac{\mathcal{P}(Y|\theta, X, f, \mathcal{M})\mathcal{P}(\theta|\mathcal{M})}{\mathcal{P}(Y|\mathcal{M})} = \alpha\mathcal{P}(Y|\theta, X, f, \mathcal{M})\mathcal{P}(\theta|\mathcal{M}) \tag{4}$$

- $\mathcal{P}(\theta|\mathcal{M})$ expresses our **prior** beliefs of the quantity of interest, θ given our assumptions dictated by $\mathcal{M}$
- $\mathcal{P}(Y|\theta, X, f, \mathcal{M})$ is the quantity we can measure directly: given values for X, θ and our functional form, what is the **likelihood** of obtaining our observable data Y
- $\mathcal{P}(\theta|X, Y, f, \mathcal{M})$ is our quantity of interest. Known as the **posterior** distribution.

Bayesian inference example

'Interesting' example, Estimating the physical quantities of galaxy clusters:

- *Observables* (X, Y) are associated with the 'boring' signals measured by a telescope
- *Parameters* of 'interest' θ are physical properties of clusters associated with these signals: mass, temperature, etc.
- We use a theory-based *generative model* f to map from observable X, and unobservable θ to our observable for Y
- n-body simulations tell us the *prior* on θ e.g. $\mathcal{P}(\theta|\mathcal{M}) = \mathcal{N}(\mu, \Sigma_P)$
- We perform a new experiment with our telescope to measure (X, Y)
- Due to thermal, CMB spectrum etc. noise, we assume our data have Gaussian errors: $\mathcal{P}(Y|\theta, X, f, \mathcal{M}) = \mathcal{N}(f(X;\theta), \Sigma_L)$

Maximum posterior estimation

- **Sampling** $\mathcal{P}(\theta|X, Y, f, \mathcal{M})$ is computationally expensive
- The "poor man's" attempt at Bayesian inference is to find the **maximum** of $\mathcal{P}(\theta|X, Y, f, \mathcal{M})$

Definition (Maximum posterior estimation)

This is unsurprisingly known as **maximum posterior estimation** (MAP, c.f. maximum likelihood estimation, MLE):

$$\max_{\theta} \mathcal{P}(\theta|X, Y, f, \mathcal{M}) \tag{5}$$

Note that:

$$\max_{\theta} \mathcal{P}(\theta|X, Y, f, \mathcal{M}) = \min_{\theta} \left[-\log(\mathcal{P}(Y|\theta, X, f, \mathcal{M})) - \log(\mathcal{P}(\theta|\mathcal{M}))\right] \tag{6}$$

The bridge

Let's compare our **optimisation** equations from the supervised machine learning and Bayesian inference cases:

$$\min_{\theta} \tau_L \mathcal{L}(f(X;\theta), Y) + \tau_R \mathcal{R}(\theta)$$
$$\min_{\theta} -\log(\mathcal{P}(Y|\theta, X, f, \mathcal{M})) - \log(\mathcal{P}(\theta|\mathcal{M}))$$

Thus by solving a supervised learning problem, we are in fact obtaining a MAP estimate where:

$$\tau_L \mathcal{L}(f(X;\theta), Y) = -\log(\mathcal{P}(Y|\theta, X, f, \mathcal{M})) \quad (7)$$
$$\tau_R \mathcal{R}(\theta) = -\log(\mathcal{P}(\theta|\mathcal{M})) \quad (8)$$

Phew... we are doing statistics afterall!

MAP estimate example

Going back to our house prediction example in the supervised learning case. We find that this is *equivalent* to obtaining a MAP estimate with:

- A three-dimensional Gaussian prior on θ with $\mu = 0$ and $\Sigma_P = \frac{1}{2\tau_R} \times \mathbb{I}$

$$\mathcal{P}(\theta|\mathcal{M}) \propto \exp\left(-\frac{2\tau_R}{2}\theta^\top \mathbb{I}^{-1}\theta\right) \Rightarrow -\log\left(\mathcal{P}(\theta|\mathcal{M})\right) \propto \tau_R \theta^\top \theta = \tau_R R(\theta) \tag{9}$$

- A ten-dimensional Gaussian likelihood on Y with $\mu = f(X;\theta)$ and $\Sigma_L = \frac{1}{2\tau_L} \times \mathbb{I}$:

$$\begin{aligned} &\mathcal{P}(Y|\theta, X, f, \mathcal{M}) \propto \exp\left(-\frac{2\tau_L}{2}(Y - f(X;\theta))^\top \mathbb{I}^{-1}(Y - f(X;\theta))\right) \\ &\Rightarrow -\log\left(\mathcal{P}(Y|\theta, X, f, \mathcal{M})\right) \propto \tau_L (Y - f(X;\theta))^\top (Y - f(X;\theta)) \\ &= \tau_L \mathcal{L}(f(X;\theta), Y) \end{aligned} \tag{10}$$

MAP estimate example conclusions

- Note we have 'absorbed' the normalisation constants associated with the prior and likelihood functions into the normalisation constant α introduced in (4)
- In our supervised learning example, we are solving a MAP problem with a Gaussian likelihood with variance $\propto 1/\tau_L$, and a Gaussian prior with zero mean and variance $\propto 1/\tau_R$
- The functional forms of $\mathcal{L}$ and $\mathcal{R}$ aren't picked out of thin air
- The regularisation constants are intricately linked to the variances of said probability distributions. Ignore them at your own peril...!
- In Bayesian inference τ_L, τ_R can be treated as random variables using *hierarchical Bayesian inference*

Summary

- We first introduced a simple theory underlying supervised machine learning (SML)
- Second we introduced the theory underpinning Bayesian inference
- Next we defined what maximum posterior estimation (MAP) is
- We then showed the SML and MAP theoretical equivalence
- Finally, for the SML example presented we considered the MAP equivalent, and showed their equivalence

Cheers for listening