Machine Learning I Lecture IV: Bayesics of Inference

Jakob H Macke
Max Planck Institute for Biological Cybernetics
Bernstein Center for Computational Neuroscience

XY.XY.2012

Plan for today

Bayesian Inference

Example: Bayesian coin toss

Conjugate priors and the exponential family

'Bayesian Inference' refers to computation of the posterior distribution over parameters given the data.

- Data $\mathcal{D} = \{t_1, t_2, \ldots, t_N\}$
- Supervised learning: $\mathcal{D} = \{(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)\}$
- Likelihood function $P(t|w)$ (parameterized by w)
- $P(D|w) = \prod_{n=1}^{N} P(t_n|w)$
- Prior distribution $\pi(w)$
- Bayes rule: Posterior $\propto$ Likelihood $\times$ Prior

$$P(w|D) = \frac{1}{Z} P(D|w)P(w) \tag{1}$$

- Probabilities must normalize to 1:

$$Z = P(D) = \int P(D|w)P(w)dw \tag{2}$$

We can use the posterior distribution to make predictions, decisions or scientific statements.

- Predictions: Predictive distribution

$$P(t^*|x^*, D) = \int_w P(t^*|x^*, w)P(w|D)dw \tag{3}$$

- Making decisions: Suppose we have calculated $P(t^*|x^*, D)$, and someone asks us to give a guess $\hat{t}$ of t^*. While we could just take the *most likely value* of t^*, it really depends on the cost function. Are mistakes in one direction as costly as mistakes in the other direction? For example, what is the cost of a false positive or false negative? Given a cost function $C(\hat{t}, t)$, one can calculate the 'Bayes-optimal' decision from the posterior distribution.
- Scientific statements: e.g. 'After observing 100 data points, we were 90% sure that the parameter θ is between -.1 and .3. Now that we have observed another 200, we are 97% sure.'

In most cases, the posterior distribution can not be calculated exactly, and approximations have to be used.

- Ignore prior, maximize likelihood $P(D|w)$: Maximum likelihood learning
- Only search for mode of posterior (Maximum a posteriori, MAP), i.e. $\text{argmax}_w P(w|D)$. In practice, maximize log-posterior.
  Q: Why is finding the mode of the posterior so much easier than finding the full posterior?
  Q: When is MAP a really bad idea?
- Use simplified model to approximate posterior: Find parameters of model $q(w, \Phi)$ such that $q(w) \approx P(w|D)$.
  - Examples: Varational Inference, Expectation Propagation, Laplace Approximation
  - Very often, a Normal approximation is used: $q(w) \approx \mathcal{N}(w|\mu, \Sigma)$.
- Use MCMC sampling to generate samples from posterior distribution

Example: Bayesian coin toss

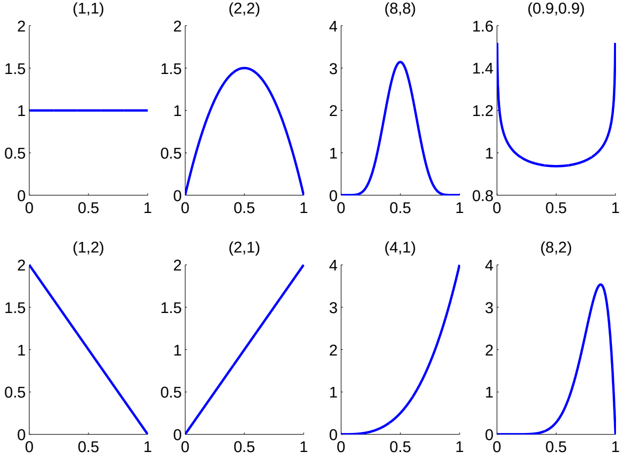
- Suppose we have N throws of a coin, $D = \{t_1, t_2, \ldots, t_N\}$
- We write $T_n = 1$ if the n-th throw was head, and $t_n = 0$ if it was tail.
- One parameter: $q \in [0, 1]$, the probability of obtaining heads
- Likelihood of one throw:

$$P(T_n = 1|q) = q \tag{4}$$
$$P(T_n = 0|q) = (1 - q) \tag{5}$$

- Likelihood of data D: [on board]

We will use a beta distribution as a prior for q.

The shape of the distribution is determined by two parameters.



We will use a beta distribution as a prior for q.

- Beta distribution:

$$\pi(q|\alpha,\alpha_2) = \frac{1}{Z} q^{\alpha_1 - 1}(1-q)^{\alpha_2 - 1} \tag{6}$$

- Normalizing constant: the 'beta function'

$$Z = \int_0^1 q^{\alpha_1 - 1}(1-q)^{\alpha_2 - 1} dq =: B(\alpha_1, \alpha_2) \tag{7}$$

- Mean and Variance:

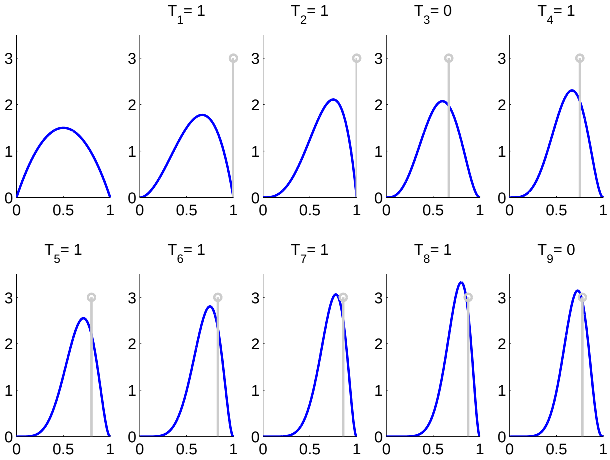
$$\mathrm{E}(q|\alpha_1,\alpha_2) = \frac{\alpha_1}{\alpha_1 + \alpha_2} \tag{8}$$

$$\mathrm{Var}(q|\alpha_1,\alpha_2) = \frac{\alpha_1 \alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)} \tag{9}$$

- Symmetric case and heuristics: [on board]

Illustration: The posterior gets more peaked as more data is coming in.

Data: $D = \{110111110\}$



The posterior distribution can be calculated in closed form.

We use S_n to denote the number of heads on the first n trials.

Maximum likelihood estimation:

[on board]

Posterior distribution:

[on board]

We can either take all the data and calculate the posterior at once, or do it sequentially as new data comes in:

[on board]

We can use the posterior distribution for predictions ...

After observing N coin-flips, what is our prediction for the next coin flip?

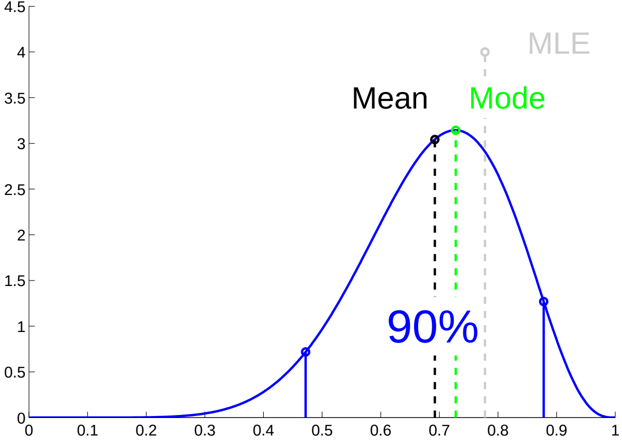
$$\begin{aligned}
P(T^* = 1|D) &= \int_0^1 P(T^* = 1|q)P(q|D)dq && (10)\\
&= \int_0^1 qP(q|D)dq && (11)\\
&= \mathrm{E}(q|D) && (12)\\
&= \frac{\alpha_1 + S_N}{(\alpha_1 + S_N) + (\alpha_2 + N - S_N)} && (13)\\
&= \frac{\alpha_1 + S_N}{(\alpha_1 + \alpha_2 + N)} && (14)
\end{aligned}$$

In our example, $\mathrm{E}(q|D) = 0.69$, and MLE $= 0.78$.

Q: What happens if N gets very large?

Note: This is a bit of a special case– in general, the predictive distribution is not simply the likelihood evaluated at the mean!!!

... for statistical reasoning ....



... and for making decisions.

Someone offers you the bet that you get 1 euros if you predict the next coin toss correctly, but you have to pay 2 euros if if you are wrong. Should you take the bet? What should you predict?

$$C(t^*, \hat{t}) = \begin{cases} -1 & \text{if } t^* = \hat{t} \\ 2 & \text{if } t^* \neq \hat{t} \end{cases} \tag{15}$$

Expected cost:

$$E(C) = \sum_{t^*=0}^{1} C(t^*, \hat{t}) P(t^*|D) \tag{16}$$

If we predict tail ($t^* = 0$):

$$E(C) = P(t^* = 1|D)C(0,1) + P(t^* = 0|D)C(0,0) \tag{17}$$

$$= 0.69(2) + 0.31(-1) = 1.07 \tag{18}$$

If we predict head ($t^* = 1$):

$$E(C) = 0.69(-1) + 0.31(2) = -0.07; \tag{19}$$

Why was inference so easy here?

- Posterior distribution had a closed form solution.
- In fact, the posterior had the same functional form as the prior, just different parameters.
- Parameters of posterior could be calculated by simply adding observations to prior parameters.
- We used a likelihood from the exponential family and its conjugate prior. In this case, Bayesian inference is always easy.
- For this reason, exponential families and conjugate priors are used extensively in Bayesian modelling, often as 'building blocks' of more complicated models.

Inference is easy whenever the likelihood is in the exponential family and the prior is its conjugate.

- Exponential family distributions have the form

$$P(\mathbf{x}|\theta) = g(\theta) f(\mathbf{x}) \exp\left(\phi(\theta)^\top S(\mathbf{x})\right) \tag{20}$$

- The conjugate prior is

$$\pi(\theta) = F(\tau, \nu) g(\theta)^\nu \exp(\phi(\theta)^\top \tau) \tag{21}$$

- Calculating the posterior:

$$\text{[on board]} \tag{22}$$

Inference is easy whenever the likelihood is in the exponential family and the prior is its conjugate.

The posterior given an exponential family likelihood and conjugate prior is

$$P(\theta|D) = F(\tau + \sum_i S(x_i), \nu + N) g(\theta)^{\nu+N} \exp\left(\phi(\theta)^\top (\tau + \sum_i S(x_i))\right) \tag{23}$$

- $\phi(\theta)$ is the vector of natural parameters
- $\sum_i S(x_i)$ is the vector of sufficient statistics
- τ are pseudo-observations
- ν is the scale of the prior

The exponential family includes most common distributions, including the Normal, Exponential, Gamma, Chi-square, Beta, Dirichlet, Bernoulli, Poisson, Wishart and the Inverse Wishart.

How can we put our coin-example into this framework?

[on board]