THE USE OF THE BAYESIAN APPROACH IN STATISTICAL LEARNING

D.M. Titterington
Department of Statistics
University of Glasgow
Glasgow, G12 QQ
Scotland, UK

SUMMARY

The Bayesian approach

Computational complications

Simulation methods (MCMC)

Variational approximations

Towards the Support Vector Machine

Bayesian approach to SVM

The Relevance Vector Machine

A (VERY) SIMPLE EXAMPLE

Data : $\mathbf{x} = (x_1, \ldots, x_n)$, a set of male birth-weights.

Model : Underlying distribution is $N(w, \sigma_0^2)$, where σ_0^2 is known.

Objective : Estimate w.

Key quantity : the **Likelihood function**

$$\text{Lik}(w; \mathbf{x}) = p(\mathbf{x}|w) \propto \exp\{-\frac{n}{2\sigma_0^2}(\bar{x} - w)^2\},$$

where $\bar{x}$ is the sample mean.

Point estimate : $\widehat{w}_{ML}$ to max $\text{Lik}(w; \mathbf{x})$;
here $\widehat{w}_{ML} = \bar{x}$.

Interval estimate : 95% Confidence Interval given by

$$\bar{x} \pm 1.96\sigma_0/\sqrt{n}.$$

Interpretation : the long-run (*over many datasets*) chance that such a C.I. contains the true w is 0.95.

Popular misinterpretation : a probability of 0.95 can be attributed to the true w being covered by the interval given *this particular dataset*.

The **Bayesian approach** provides interval estimates that do allow this interpretation, by creating a density $p(w|\mathbf{x})$.

The model and likelihood provide a $p(\mathbf{x}|w)$. How to turn this around? Use **Bayes' Theorem**:

$$\begin{aligned} p(w|\mathbf{x}) &= \frac{p(\mathbf{x}|w)p(w)}{p(\mathbf{x})} \\ &\propto p(\mathbf{x}|w)p(w) \end{aligned}$$

What is $p(w)$? Called the **prior** density for w - 'before' the data - whereas $p(w|\mathbf{x})$ is called the **posterior** for w - 'after' the data.

Advantages :

1. quantifiable prior knowledge can be incorporated;

2. different analysts provide different inferences.

Problematic issues :

1. we are combining two sorts of densities;

2. different analysts provide different inferences;

3. (possible computational problems).

BIRTHWEIGHTS EXAMPLE

Suppose, for prior, $w \sim N(a, b^2)$. Then

$$\begin{aligned} p(w|\mathbf{x}) &\propto \exp\{-\frac{n}{2\sigma_0^2}(\bar{x}-w)^2 - \frac{1}{2b^2}(w-a)^2\} \\ &\propto \exp\{-\frac{1}{2B^2}(w-A)^2\}, \end{aligned}$$

where

$$\begin{aligned} A &= \frac{\frac{n}{\sigma_0^2}\bar{x} + \frac{a}{b^2}}{\frac{n}{\sigma_0^2} + \frac{1}{b^2}} \\ \frac{1}{B^2} &= \frac{n}{\sigma_0^2} + \frac{1}{b^2}. \end{aligned}$$

Therefore, $w|\mathbf{x} \sim N(A, B^2)$.

Point estimate : $\widehat{w} = A$.

95% Interval Est. : $A \pm 1.96B$.

(If $b \to \infty$ then $\widehat{w} \to \bar{x}$, I.E. $\to$ C.I!)

MALE AND FEMALE BIRTHWEIGHTS

Suppose n birthweights $\mathbf{x}$ are recorded from a **mixture** of males and females but nobody notes which babies were males and which were females; i.e. for each baby the sex-indicator is *missing*.

Assume that the male and female birthweight distributions are Gaussian, with the same *known* variance σ_0^2 but with unknown and possibly different means w_M and w_F. Also, *assume* that the proportions of males and females in the population are equal. Then

$$p(\mathbf{x}|\mathbf{w}) \propto \prod_i [\frac{1}{2}\text{exp}\{-\frac{1}{2\sigma_0^2}(x_i - w_M)^2\} + \frac{1}{2}\text{exp}\{-\frac{1}{2\sigma_0^2}(x_i - w_F)^2\}].$$

We might assume that, for priors,

$$p(\mathbf{w}) = p_M(w_M)p_F(w_F),$$

where each of the factors on the RHS is a Gaussian pdf.

In this case $p(\mathbf{w}|\mathbf{x})$ is not simple, and calculation of point and interval estimates is harder, a common feature of contexts with *incomplete data*.

(Note: the same is true for non-Bayesian inference - explicit formulae for ML estimates are not available.)

What to do?

Maximum likelihood : use an iterative algorithm.

Bayesian approach : for point estimates, e.g. posterior modes, do as for ML; for other purposes 'approximate' $p(\mathbf{w}|\mathbf{x})$ either through some deterministic approximation or by simulating a large number of realisations from $p(\mathbf{w}|\mathbf{x})$.

MORE ON THE BAYESIAN APPROACH:

Introduce missing data indicators

$$\mathbf{z} = (z_1, \ldots, z_n),$$

where $z_i = 1$ if male and $z_i = 0$ if female. Then

$$p(\mathbf{x}, \mathbf{z}|\mathbf{w}) \propto \prod_i [\exp\{-\frac{1}{2\sigma_0^2}(x_i - w_M)^2\}]^{z_i} \times [\exp\{-\frac{1}{2\sigma_0^2}(x_i - w_F)^2\}]^{(1-z_i)}.$$

If the sex-indicators are known and independent Gaussian priors are assumed for w_M and w_F then the posterior densities are also independent and Gaussian, with (hyper)parameters that can be determined exactly.

Simulation approach: choose an initial $\mathbf{z}$ and iteratively simulate $\mathbf{w}$ and $\mathbf{z}$ from their full conditional densities. The resulting equilibrium distribution is the joint posterior for $\mathbf{w}$ and $\mathbf{z}$ (*Gibbs sampling*).

Deterministic approach: propose a simple form for the joint posterior for $\mathbf{w}$ and $\mathbf{z}$ and optimise within that form, providing a so-called *variational approximation* for the joint posterior of $\mathbf{w}$ and $\mathbf{z}$, from which the marginal posterior for $\mathbf{w}$ is usually easily obtained.

MORE ON THE VARIATIONAL APPROXIMATION

Suppose $Q(\mathbf{w}, \mathbf{z})$ defines an approximation to $p(\mathbf{w}, \mathbf{z}|\mathbf{x})$ and suppose we *propose* that Q takes the factorised form

$$Q(\mathbf{w}, \mathbf{z}) = Q_{w_M}(w_M) Q_{w_F}(w_F) \prod_i Q_{z_i}(z_i),$$

where the factors are chosen to optimise

$$\mathrm{KL}(Q, p) = \int_{\mathbf{w}} \sum_{\mathbf{z}} Q \ \log \ (Q/p),$$

the Kullback-Leibler Directed Divergence. If independent Gaussian priors are chosen for w_M and w_F then Q_{w_M} and Q_{w_F} are Gaussian, with hyperparameters obtained from nonlinear equations.

This is a standard pattern for these variational approximations.

TOWARDS THE SUPPORT VECTOR MACHINE

A REGRESSION MODEL

$$y_i = f(x_i) + \eta_i,$$

for $i = 1, \ldots, n$, where y is the *response*, f is the *regression function*, x are *covariates* and η is *noise*. Propose a formulation in which

$$f(x) = h(x)^T w + w_0,$$

where $h(x)$ is a vector of *basis functions*.

How to choose/estimate (w, w_0)? Define a *regularised risk function*

$$R(w, w_0) = \sum_i \Delta\{y_i - h(x_i)^T w - w_0\} + \frac{\lambda}{2} w^T w,$$

where Δ is a *loss function* and λ is a *regularisation parameter* or *tuning constant*.

Point estimation: choose $(\hat{w}, \hat{w}_0)$ to min $R(w, w_0)$.

Examples include *linear regression* ($\Delta(u) = u^2, \lambda = 0$), *ridge regression* ($\Delta(u) = u^2, \lambda \geq 0$), *robust estimation*, *spline smoothing* and ...

SUPPORT VECTOR MACHINES (SVM)

Here

$$\begin{aligned} \Delta(u) &= 0 \quad &\text{for} \quad |u| < \varepsilon \\ &= |u| - \varepsilon \quad &\text{for} \quad |u| \geq \varepsilon \end{aligned}$$

Minimisation of $R(w, w_0)$ is explicit if $\Delta(u) = u^2$ and requires quadratic programming in the SVM. The SVM solution takes the form

$$\widehat{w} = \sum_i \alpha_i h(x_i),$$

for certain $\{\alpha_i\}$, many of which turn out to be zero. The data points with nonzero α_i are the *support vectors*. Also

$$f(x) = \sum_i \alpha_i h(x_i)^T h(x) + \widehat{w}_0,$$

with $\widehat{w}_0$ obtained from any SV.

BAYESIAN APPROACH

Interpret $R(w, w_0)$ as -log $p(w, w_0|D)$, where D denotes the data, $\{(y_i, x_i), i = \ldots, n\}$. Thus, we interpret $-\frac{\lambda}{2}w^T w$ as -log $p(w, w_0)$ and $\sum_i \Delta\{y_i - h(x_i)^T w - w_0\}$ as -log $p(D|w, w_0)$.

Clearly, the $(\widehat{w}, \widehat{w}_0)$ that minimise $R(w, w_0)$ can be interpreted as the *posterior mode*.

BAYESIAN INTERVAL ESTIMATES

If $R(w, w_0)$ *is quadratic* then the equivalent $p(w, w_0|D)$ is Gaussian and interval estimation is quite easy.

If $R(w, w_0)$ *is not quadratic*, use simulation or *Laplace approximation*, a Gaussian approx based on quadratic Taylor expansion of $R(w, w_0)$ about $(\widehat{w}, \widehat{w}_0)$.

WHAT ABOUT λ?

λ is a hyperparameter. A full Bayesian approach puts a hyperprior on λ, but in many contexts a value is ‘plugged in’, e.g. as follows.

Write the prior for (w, w_0) as $p(w, w_0|\lambda)$ and consider

$$p(D|\lambda) = \int p(D|w, w_0)p(w, w_0|\lambda)\ dw\ dw_0,$$

called the *marginal likelihood* or the *Type II likelihood* or the *evidence*. Choose for λ the maximiser $\widehat{\lambda}$ of $p(D|\lambda)$.

Calculation of the integral: easy if the integrand corresponds to a Gaussian density for (w, w_0), but even then maximisation wrt λ is non-explicit; otherwise, use Laplace approximation to create a Gaussian integrand or use a variational approximation.

THE RELEVANCE VECTOR MACHINE (RVM) (Tipping, 2000)

Write the model as

$$y = \sum_i \alpha_i h(x_i)^T h(x) + \alpha_0 + \eta.$$

Choose Δ quadratic, corresponding to Gaussian noise, and let τ be the inverse of the variance of the noise. *Consequence*: $p(\alpha|D)$ is Gaussian.

Crucial modification: in the prior, assume

$$\alpha \sim N\{0, \text{diag}(\lambda_0, \lambda_1, \ldots, \lambda_n)\},$$

so that there is a hyperparameter for each data-point.

The marginal likelihood $p(D|\{\lambda_i\}, \tau)$ can be calculated explicitly, but it involves inverting an $n \times n$ matrix, not to mention numerical optimisation. However empirical work by Tipping shows that many of the λ_i's get very large, so that the resulting data-point will not be a support vector; typically, the number of support vectors with the RVM is much less than with SVM, without degradation in performance.

Bishop and Tipping (2000) use variational approximation for 'calculating' the marginal likelihood.

SOME EMPIRICAL RESULTS (Tipping)

REGRESSION

Support points for SVM/RVM

Dataset	n	SVM	RVM
1	240	116	59
2	240	110	7
3	240	106	12
4	481	143	39

CLASSIFICATION

Support points for SVM/RVM

Dataset	n	SVM	RVM
Pima Indians	200	109	4
USPS Digits	7291	2540	316

COMMENTS COMPARING THE MCMC AND DETERMINISTIC APPROACHES

In principle the MCMC is 'exact' given enough computing power.

In principle the deterministic approaches are not exact.

However the MCMC approach may be prohibitive in very large-scale problems, and the deterministic approximations may be adequate in practice.

Also, if there are large amounts of data, the deterministic approaches may provide 'asymptotically respectable' approximations - research on this is in progress!

REFERENCES

BISHOP, C.M. and TIPPING, M.E. (2000). Variational relevance vector machines. In *16th Conf. Uncertainty in Artificial Intelligence* (C. Boutilier and M. Goldszmidt, eds.) 46–53. Morgan Kaufmann, San Mateo, CA.

BURGES, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining & Knowledge Discovery* **2**, 121–167.

CHU, W., KEERTHI, S.S. and ONG, C.J. (2001). Bayesian inference in support vector regression. Technical Report CD-01-15, Nat. Univ. Singapore.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer, New York.

KWOK, J. T.-Y. (2000). The evidence framework applied to support vector machines. *IEEE Trans. Neural Networks* **11** 1162–1173.

ROBERT, C.P. (1992?). *L'Analyse Statistique Bayesienne.* Economica, Paris.

SEEGER, M. (2000). Bayesian model selection for support vector machines, Gaussian processes, and other kernel classifiers. In *Advances in Neural Information Processing, Vol.12* (S.A. Solla, T.K. Leen and K.-R. Müller, eds.). MIT Press, Cambridge, MA.

TIPPING, M.E. (2000). The relevance vector machine. In *Advances in Neural Information Processing, Vol.12* (S.A. Solla, T.K. Leen and K.-R. Müller, eds.). MIT Press, Cambridge, MA.

TITTERINGTON, D.M. (2003). Bayesian methods for neural networks and related models. Submitted for publication.