Bayesian regression:

Esmat Hashemi ,Asieh Gholami

s_hashemi1368@yahoo.com,Gholami19@yahoo.com

Department of statistics, payame nor university, Tehran, Iran

Abstract

We are to estimate a point of the W parameter vectors in classic regression models. In contrast, we define uncertainty in W in a Bayesian approach by p(w) probability distribution. The observations from data points change through the Bayesian theorem by the data which are intermediate by probability function in likelihood.

Keyword : Oval-regression- functions

Introduction:

Previous distribution of p (w) indicates uncertainty in w by taking into consideration all data and without losing the generality we can write:

1 – 1 : $p(\mathrm{w}|\alpha) \propto \exp\{-\alpha\,\Omega(\mathrm{w})\}$

We choose special examples. The Gaussian distribution is for $p(w|\alpha)$

As follows:

1-2: $p(\mathrm{w}|\alpha) = (\frac{\alpha}{2\pi})^{\frac{M}{2}} \exp\left\{-\frac{\alpha}{2}||\mathrm{w}||^2\right\}$

Now we can use Bayesian theories to state distribution like a priori distribution and posteriori function for w.

Suppose that we want to use previous distribution to find a point estimate for w which is to maximize a posteriori distribution or at least equal to minimize negative logarithm of distribution:

1-3: $p(w|t,\alpha,\sigma^2) \propto p(w|\alpha)L(w)$

Where: L (w) = $p(w|t, \alpha, \sigma^2)$

Suppose we want to use a priori distribution to find a point estimate for w which is to maximize a posteriori distribution or equivalently minimize the negative logarithm of distribution, then equal a posteriori distribution logarithm maximization is to minimize.

1 – 4: $\frac{1}{2\sigma^2}\sum_{n=1}^{N}|y(X_n; w) - t_n|^2 + \frac{\alpha}{2}\Omega(w)$

So we see there is close similarity between such Bayesian and ordinary views on the basis of minimizing error function and the latter was gained by a special approximation on the Bayesian approach; particularly if it is a new value of x, the a priori distribution t is the total sum and having set aside the product from the probability orders on w.

1-5: $p(t|t\,,\alpha\,,\beta) = \int p(w\,|t\,,\alpha\,,\sigma^2)\;\; p(t|w,\sigma^2\,)\,dw$

The appropriate variable is unknown in most of the applications though sometimes σ^2 is clear.

1 – 2: Support vector machine (SVM):

SVM is one of the methods to learn under supervision to be used for stratification and regression; recently it has had a good efficiency compared to previous methods to stratify, for example, perception nervous networks. The SVM function is based on grouping data linearly and in dividing linearly data we try to select a line with more confidence margin. The SVM algorithm solution is grouped as the pattern distinction algorithms. In SVM algorithm whenever it is necessary to distinguish the pattern or group things in special classes it is possible to use it; we provide the pattern matrix and then select the kernel function and then select kernel function parameter and c value in order to use SVM . SVM has been usually defined clearly in such form and leads to prediction based on the function though the application has not been defined as ordinary.

1 – 6: y(x, w) = $\sum \omega_i$k(x, x_i) +ω_0

Its objective function is to minimize the error measurement in the set and the same time, maximize the margin between two classes. We need p(t|z) distribution for Bayesian activities from logarithm applied programs; such distribution indicates uncertainty in prediction and shows several advantages

such as favorable and flexible decision as an output or other sources of probable data.

1-3: Special models:

We suppose a set of input indicators $\{a_{nd}, t_{an}\}$ where there are a model of turbulence with zero average and σ^2 variance. So:

1 – 7: $p(t_n|X) = \mathcal{N}(t_n|y(X_n; w), \sigma^2)$

Where the y(x, w) function is defined in (1-6). By virtue of tan independence the probability of complete data set may be written as follows:

1-8: $p(t|w, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2}\|t - \Phi w\|^2\right\}$

The matrix: $\Phi = [\phi(X_1), \phi(X_2), \dots, \phi(X_N)]^T$ is known as the designer matrix.

$\phi(X_n) = [1, K(X_n, X_1), K(X_n, X_2), \dots, K(X_n, X_N)]^T$

And

$t = (t_1, \dots, t_N)^T$ و $w = (w_0, \dots, w_1)^T$

As many parameters in this model, for example, as the test sample we expect the likelihood maximum method leads to decline to more than the connections. We use Gaussian a priori distribution to complete the characters of previous hierarchy.

1 – 9: $p(\omega|\alpha) = \prod_{i=0}^{N} \mathcal{N}(\omega_i|0, \alpha_i^{-1})$

These quantities are examples of scale parameter and convenient references were given by gamma for them.

1 – 10: $p(\alpha) = \prod_{i=0}^{N} Gamma\ (\alpha_i|a, b)$

1 – 11: $p(\beta) = Gamma\ (\beta|c, d)$

Where: $Gamma(\alpha|a, b) = \Gamma(a)^{-1} b^a \alpha^{a-1} e^{-b\alpha}$ is gamma function?

Where the gamma reference does not contain useful data in the level of $a \rightarrow 0$ and b→0. The ultra parameters may be inconvenient. We may solve this problem by little values: a=b=c=d=10^{-4} .

Where:

We consider monotone previous scale is with a=b=c=d=0 . The Bayesian inference is defined by computations:

1 – 12: $p(w,\alpha,\sigma^2|t) = \frac{p(t|w,\alpha,\sigma^2)p(w,\alpha,\sigma^2)}{p(t)}$

Then by virtue of a new test point, x_* is the prediction for the goal t_* in previous defined distribution conditions.

1-13: $p(t_*|t) = \int p(t_*|w,\alpha,\sigma^2|t)p(w,\alpha,\sigma^2|t)dw\, d\alpha\, d\sigma^2$

Where we cannot compute directly $p(w,\alpha,\sigma^2|t)$ from 1-13 because we cannot do ordinary integral.

1-4: Perhaps you ask why Gaussian previous selection should be stated before any priority for dispersed models. We may integrate previous ultra-parameters in order to have such view. It is possible to integrate α independently or take into consideration correctly for previous gammas on additional parameters. It seems previous dispersed like Laplacian reference reached its peak in zero.

$p(\omega_i) \propto \exp\ (-|\omega_i|)$

Which was used to achieve dispersion in Bayesian field.

1-14: $P(w \mid t, \alpha) = (2\pi)^{-(N+1)/2}|\sum|^{-1/2} \exp\left\{-\frac{1}{2}(w-\mu)^T\sum^{-1}(w-\mu)\right\}$

This is a posteriori and mean covariance:

1-15:

$\Sigma = (\sigma^{-2}\Phi^T\Phi + A)^{-1}$ $\mu = \sigma^{-2}\sum \Phi^T t$

$A = diag(\alpha_0, \alpha_1, \ldots, \alpha)$

$\int p(t_*|\alpha,\sigma^2)p(\alpha,\sigma^2|t)d\alpha\, d\sigma^2 \simeq p(t_*|\alpha_{MP},\sigma^2_{MP})$

The symbol may be visualization of a mental test when we use the $\Phi_j(x)$ and $\Phi_i(x)$ functions in the same way. Maximum relation:

1-16: $p(\alpha, \sigma^2| t) \propto p(t|\alpha, \sigma^2)\, p(\alpha)\, p(\sigma^2)$

Considering α and β for previous ultra-parameter state depends on only logarithm margin probability in $p(t|\alpha, \sigma^2)$ we need following equation to maximize or equalize :

1-17:

$$\mathcal{L}(\alpha) = \ln p(t|\alpha,\sigma^2) = \ln \int_{-\infty}^{\infty} p(t|w,\sigma^2)p(w|\alpha)dw$$

$$= -\frac{1}{2}[N\ln 2\pi + \ln|C| + t^T C^{-1}t]$$

With:

1 – 18: $C = \sigma^2 I + \Phi A^{-1}\Phi^T$

1-5: Prediction:

In maximization the prediction is done on the basis of a posteriori distribution by maximizing for a new data x_*.

1-19: $p(t_*|t,\alpha_{MP},\sigma^2_{MP}) = \int p(t_*|w,\sigma^2_{MP})p(w|t,\alpha_{MP},\sigma^2_{MP})dw$

Considering both conditions in the Gaussian may be calculated easily we have:

1 – 20: $(t_*|t,\alpha_{MP},\sigma^2_{MP}) = \mathcal{N}(t_*|y_*,\sigma^2_*)$

With: $y_* = \mu^T\phi(x_*)$

$\sigma^2_* = \sigma^2_{MP} + \phi(x_*)^T\sum\phi(x_*)$

Marginal likelihood:

C analysis is as follows:

1 – 21:

$C = \sigma^2 I + \sum_{m\neq I}\alpha_m^{-1}\phi_m\phi_m^T + \alpha_i^{-1}\phi_i\phi_i^T = C_{-i} + \alpha_i^{-1}\phi_i\phi_i^T$

1 – 22:

$$\mathcal{L}(\alpha) = -\frac{1}{2}[N\ln(2\pi) + \ln|C_{-i}| + t^T C_{-i}^{-1}t$$

$$-\ln\alpha_i + \ln\left(\alpha_i + \phi_i^T C_{-i}^{-1}\phi_i\right) - \frac{\left(\phi_i^T C_{-i}^{-1}t\right)^2}{\alpha_i + \phi_i^T C_{-i}^{-1}\phi_i}]$$

$$= \mathcal{L}(\alpha_{-1}) + \frac{1}{2}\left[\ln\alpha_i - \ln(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i}\right]$$

$$= \mathcal{L}(\alpha_{-i}) + \ell(\alpha_i)$$

Where we use following data to simplify:

1-23: $s_i \triangleq \phi_i^T C_{-i}^{-1} \phi_i \quad , \quad q_i \triangleq \phi_i^T C_{-i}^{-1}$

$$\alpha_i = \frac{s_i^2}{q_i^2 - s_i}$$

1-24:

$\alpha_i = \infty \qquad if\ q_i^2 \leq s_i$

Here we begin with the quantity s_i and qi.

1-25: $S_i = \phi_i^T C^{-1} \phi_i \quad , \quad Q_i = \phi_i^T C^{-1} t$

Which is given as follows?

(1-26)$s_i = \frac{\alpha_i S_i}{\alpha_i - S_i} \quad , \quad q_i = \frac{\alpha_i Q_i}{\alpha_i - S_i}$

The expressions are when $q_i = Q_i$, si = S_i and $\alpha_I = \infty$.

1 – 27: $S_i = \phi_i^T B \phi_i - \phi_i^T B \Phi \hat{\sum} \Phi^T B \phi_i$

$S_i = \phi_i^T B \phi_i - \phi_i^T B \Phi \hat{\sum} \Phi^T B \phi_i$

t^ ≡ t and $\hat{\mu} \equiv \mu \ and \ B \equiv \sigma^{-2}$ and $\hat{\Sigma} \equiv \Sigma$

If Φ_I is in the model , $\alpha_i < \infty$, $q_i^2 \leq S_i$ is still possible after Φ_I .

If Φ_I is not in the model, $\alpha_I = \infty$ and $q_i^2 > S_i$ are possible, ı Φ_i may be added.

If Φ_I is in the model and $q_i^2 > S_i$, then it is possible to estimate α_I again.

All performances were done to increase the likelihood function; so we have a framework to change the model by adding and eliminating essential function in the probability rule method.

For the variance σ^2 we may we may estimate the equation gained for turbulence variable variance inference.

1-28: $(\sigma^2)^{new} = \frac{\|t - \Phi\mu\|^2}{N - M + \sum_i \alpha_i \sum_{ii}}$

The dominator N indicates some examples.

1-6: Stratification:

Bayesian stratification is to describe the framework by Bernoulli probability and link function to change in the intended algorithm; as a result, there is an additional approximation step in algorithm.

We use the link function $\sigma(y) = 1/(1+e^{-y})$ and Bernoulli distribution P (t|x) so we have the density function as follows:

1-29: $P(t|w) = \prod_{n=1}^{N} \sigma\{y(X_n; w)\}^{t_n} [1 - \sigma\{y(X_n; w)\}]^{1-t_n}$

1 – Previous distribution is as follows for present values of α:

$p(w|t, \alpha) \propto p(t|w)\, p(w|\alpha)$

The equation is to find the maximum value:

1-30: $\ln\{p(t|w)p(w|\alpha)\} =$

$$\sum_{n=1}^{N} [t_n \ln y_n + (1 - t_n) \ln(1 - y_n)] - \frac{1}{2} w^T A W$$

With:

$y_n = \sigma\{y(X_n; w)\}$

Second Newton method is to minimize the repeated squares of algorithm to find $\hat{\mu}$.

2- Laplace method is only a second grade approximation.

1-31: $\nabla_w \nabla_w \ln p(w|t, \alpha)\,|_{\hat{\mu}} = -(\Phi^T B \Phi + A)$

Where: B = di$a$g (β_0,β_1 , β_2 , … ,β_N)

And:

$\beta_n = \sigma\{y(x_n)\}[1 - \sigma\{y(x_n)\}]$

In the model p (w|t,α) we may see the local effects of classic problems around μ by:

1 – 32: $\widehat{\sum} = \left(\Phi^T B \Phi + \mathrm{A}\right)^{-1}\big)$

$\hat{\mu} = \widehat{\sum} \Phi^T \mathrm{B} \hat{t}$

1-33:

$$\hat{t} = \Phi\hat{\mu} + B^{-1}(t - \sigma\{\Phi\hat{\mu}\})$$

These equations are used to solve the least squares. We see that the Laplace approximation depends on effectively a regression with indicator $t\ and$ turbulence data in which reverse variance is presented for n by:

$$\beta_n = \sigma\{y(x_n)\}[1 - \sigma\{y(x_n)\}]$$

1-7: Obvious artificial date: Regression:

The function sinc(x) =sin(x)/X is to show the regression vector; we describe the kernel function as follows:

1-34: $K(x_m, x_n) = 1 + x_m x_n + x_m x_n \min(x_m, x_n)$

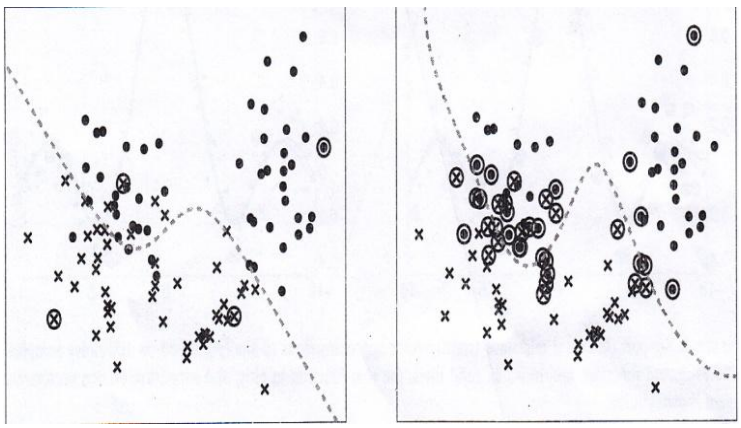
$$-\frac{x_m + x_n}{2}\min(x_m, x_n)^2 + \frac{\min(x_m, x_n)^3}{3}$$

Where the kernel function to describe a set of primitive function is:

$\Phi_n(x)=K(x_m,x_n)$, N=1,…,n

We use the artificial data produced in 2 later. The class 1 is indicated by x and class two by o in order to show graphically the connection vectors to stratify each two classes. The Bayesian error is about 8 percent by integrating Gaussian 2 mixture in the intervened stratums and the Gaussian kernel is used as follows:

1-35:$K(X_m, X_n) = \exp(-r^{-2}\|X_m - X_n\|^2)$



References:

[1] Baser. B. E., Guyton, I. M., and Vapid, V. N. A training algorithm for optimal margin

[2] Bishop, C.M.andM.E.Tipping (2000). Variationalrelevancevectormachins.In C.Boutilierand M.Goldszmidt (Ends), Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence,pp. 46-53. Morgan Kaufmann

[3] oser,B.,I.Guyon,andV.N.Vapnik(1992).Atrainingalgorithmforoptimalmarginclassifiers.In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp.144-152.

[4] Classifiers.InD.Haussler, editor, 5thAnnualACMWorkshoponCOLT, pages144 –152, Pittsburgh, PA, 1992. ACM Press

[5] Rich, S.Becker, andZ.Ghahramani (Eds.), AdvanceshnNeuralInformationProcessingSystems 14, pp.383-389. MIT Press.

[6] Faull, A. C. and M. E. Tipping (2002).Analysis of sparse Bayesian learning. In T. G.

[7] Mackay, D.J.C. (1992), Theevidenceframeworkappliedtoclassificationnetwork.NeuralComputation 4(5), 720-736.

[8] Mackay, D. J. C. (1994). Bayesian methods for back propagation network. In E. Doman, J. L.vanHemmen, andK.Schulten (Eds.), ModelsofNeural NetworksIII, Chapter6, pp.211-254.Springer.

[9] Neal, R. M. (1996). Bayesian Learning for Neural Networks. Springer.

[10] Ripley, B. D. (1996). Pattern recognition and neural networks. Cambridge University Press.

[11] Scholkopf,B.,C. J.CBburges,andA. J.Smola (Eds)(1999).Advancesin Kernel Methods:Suport Vector Learning.MIT Press

[12] TippingM. E. (2001). Sparse Bayesian learning and the relevance vector machine. Journal oflikelihoodmaximizationforsparseBayesianmodels.InB.FreyandC.M.Bishop (Eds.),Artificial Intelligence and Statistics. To appear.

[13] Tipping, M. E. (2000). The Relevance Vector Machine. IN S. A. Sololá, T. K. Len, and K>-R.Muller (Eds.), Advances in Neural Information Processing Systems 12, pp.652-658.MIT Press.

[14] Vapnik, V.N., S.E.Golowich, andA.J.Smola (1997).Supportvectormethodforfunctionapproximation, regression estimation and signal processing. In M. Composer, M. I. Jordan, and T.Pestsche (Eds.), Advances in Neural Information Processing Systems 9. MIT Press.MachineLearning

[15] Bayesian Regression and Classification.

[16] Vapid, V. N. (1998). Statistical Learning Theory. Wiley.

[17] Wickham's, P. M. (1995). Bayesian regularization and pruning using a Laplace prior. NeurelComputation 7(1), 117-143.

[18]Williams, C. K. I. (1997). Prediction with Gaussian processes: From linear regression to liner prediction and beyond. Technical report, NCRG, Aston University, Birmingham, U. K