[Paper review 1]

A Practical Bayesian Framework for Backpropagation Networks (David J.C. MacKay, 1992)

[Abstract]

How Bayesian Framework can be applied in Neural Network

- comparing "network architecture"
- "stopping rule" for network pruning
- "weight decay & regularizer"
- effective number of parameters
- quantified estimates of the error bars
- alternative learning & interpolation models

[Notation]

- w : weight (parameter)
- D : dataset $(= \{x^m, t^m\})$
- $\mathcal{A}$: network architecture
- mapping : $y(x; w, \mathcal{A})$ (= $f(x)$)

1. The Gaps in Backprop

Finding w

- 1) small error E_D (train error)

  $E_{\mathrm{D}}(D \mid \mathbf{w}, \mathcal{A}) = \sum_m \frac{1}{2}[\mathbf{y}(\mathbf{x}^m; \mathbf{w}, \mathcal{A}) - \mathbf{t}^m]^2$
- 2) generalize well (test error)

(Plain) Back propagation : perform gradient descent on E_d in w space

Modification to the plain back-prop

- add "extra regularizing terms" $E_w(w)$

  (penalize large weight)
- weight energy (= weight decay) : $E_w(w \mid A) = \sum_i \frac{1}{2} w_i^2$

Thus, target cost function to minimize :

- $M = \alpha E_{\mathrm{W}}(\mathbf{w} \mid \mathcal{A}) + \beta E_{\mathrm{D}}(D \mid \mathbf{w}, \mathcal{A})$
  - $\alpha E_{\mathrm{W}}(\mathbf{w} \mid \mathcal{A})$: regularizer
  - $\beta E_{\mathrm{D}}(D \mid \mathbf{w}, \mathcal{A})$: train loss
- gradient descent of M treats all data points equally

1-1. What is Lacking

Popular ways of comparing networks with different parameters

- ex 1) test on "Unseen dataset"
  - problem : large test set may be needed to reduce the signal-to-noise ratio in the test
- ex 2) Cross-validation
  - problem : computationally demanding

∴ need objective criteria for setting free parameters & comparing alternative solutions, that depend ONLY on the TRAIN DATASET

This paper fills the holes in the NN(=Neural Network)

- 1) objective criteria for comparing NN with different $\mathcal{A}$ (Network architecture)
  - given one $\mathcal{A}$, there may be more than one minimum of objective function M
- 2) objective criteria for setting "decay rate" (= α)
- 3) objective choice of "regularizing function" (= E_w)
- 4) objective criteria for choosing NN vs different model (ex. splines, radial basis functions)

1-2. The Probability Connection

Probabilistic view to solve the problems above

- Likelihood : $P(\mathrm{t}^m \mid \mathbf{x}^m, \mathbf{w}, \beta, \mathcal{A}) = \frac{\exp[-\beta E(\mathrm{t}^m \mid \mathbf{x}^m, \mathbf{w}, \mathcal{A})]}{Z_m(\beta)}$,

  where $Z_m(\beta) = \int dt \exp(-\beta E)$
  - β is a measure of the presumed noise included in t
  - E is error

    (if E is quadratic error function = t includes additive Gaussian noise with variance $\sigma_v^2 = 1/\beta$)

- Prior : $P(\mathbf{w} \mid \alpha, \mathcal{A}, \mathcal{R}) = \frac{\exp[-\alpha E_{\mathrm{W}}(\mathbf{w} \mid \mathcal{A})]}{Z_{\mathrm{W}}(\alpha)}$

  where $Z_w(\alpha) = \int d^k \exp(-\alpha E_w)$
  - α is a measure of the characteristic expected connection magnitute
  - if E_w is quadratic function = w are expected to come from Gaussian with "zero mean" & "variance $\sigma_w^2 = 1/\alpha$)

- Posterior Probability : $P(\mathbf{w} \mid D, \alpha, \beta, \mathcal{A}, \mathcal{R}) = \frac{\exp(-\alpha E_{\mathrm{W}} - \beta E_{\mathrm{D}})}{Z_{\mathrm{M}}(\alpha, \beta)}$

  where $Z_{\mathrm{M}}(\alpha, \beta) = \int d^k \mathbf{w} \exp(-\alpha E_{\mathrm{W}} - \beta E_{\mathrm{D}})$

Under this framework,

- minimization of $M = \alpha E_w + \beta E_D$ is finding the most probable parameters w_{MP}
- backprop's energy functions E_D and E_W, and to parameters α and β
- makes it possible to predict "average generalization ability" of NN

how to estimate "Saliency" of a weight = change in M when the weight is deleted (Le Cun et al (1990))

Hessian of M can be used to assign error bars to the parameters (Denker and Le Cun (1991))

2. Review of Bayesian Regularization and Model Comparison

How the framework can be set to handle NN, where "the landscape of $M(w)$ is certainly not quadratic"

2-1. Determination of α and β

posterior for α and β

- $P(\alpha, \beta \mid D, \mathcal{A}, \mathcal{R}) = \frac{P(D|\alpha,\beta,\mathcal{A},\mathcal{R})P(\alpha,\beta)}{P(D|\mathcal{A},\mathcal{R})}$

Evidence for α and β (assume uniform prior)

- $P(D \mid \alpha, \beta, \mathcal{A}, \mathcal{R}) = \frac{Z_{\mathrm{M}}(\alpha,\beta)}{Z_{\mathrm{W}}(\alpha) Z_{\mathrm{D}}(\beta)}$

  where $Z_D = \int d^N D e^{-\beta E_D}$

Example)

Setting :

- $E_{\mathrm{D}}(D \mid \mathbf{w}, \mathcal{A}) = \sum_m \frac{1}{2}[\mathbf{y}(\mathbf{x}^m; \mathbf{w}, \mathcal{A}) - \mathbf{t}^m]^2$
- $E_w(w \mid A) = \sum_i \frac{1}{2} w_i^2$
- N : number of dataset
- k : dimension of w (number of free parameters)

Then,

- $Z_D = (\frac{2\pi}{\beta})^{N/2}$
- $Z_W = (\frac{2\pi}{\alpha})^{k/2}$

- $Z_{\mathrm{M}} \simeq e^{-M(\mathrm{w}_{\mathrm{MP}})}(2\pi)^{k/2}\det{}^{-1/2}A$

  (where $\mathbf{A} = \nabla\nabla M$ is the Hessian of M evaluated at w_{MP})

The maximum of evidence (=$P(D \mid \alpha, \beta, \mathcal{A}, \mathcal{R})$) has following properties

- $\chi^2_W \equiv 2\alpha E_W = \gamma$
- $\chi^2_{\mathrm{D}} \equiv 2\beta E_{\mathrm{D}} = N - \gamma$

where, $\gamma = \sum_{a=1}^{k} \frac{\lambda_a}{\lambda_a + \alpha}$ (λ_a : eigenvalues of the quadratic form βE_D in the natural basis of E_W)

2-2. Comparison of Different Models

- simply evaluate the evidence (= $P(D \mid \mathcal{A}, \mathcal{R})$)
- integrate the evidence w.r.t (α, β)

  $P(D \mid \mathcal{A}, \mathcal{R}) = \int P(D \mid \alpha, \beta, \mathcal{A}, \mathcal{R}) P(\alpha, \beta) d\alpha d\beta$

3. Adapting the Framework

M has many local minima .

But it's ok, since "we need to evaluate Z_M, not M" to evaluate the evidence for all the candidate models.

(Z_M is the numerator part of evidence)

how to find local $Z_{\mathrm{M}} \simeq e^{-M(\mathrm{w}_{\mathrm{MP}})}(2\pi)^{k/2}\det{}^{-1/2}A$?

- need to evaluate (or approximate) the inverse Hessian of M

  need to evaluate (or approximate) its determinant and/or trace
- solved by Denker and Le Cun (1991)

4. Demonstration

4-1. Relation to Generalization Error

How good a predictor of network quality the EVIDENCE is

- need to check "relationship between EVIDENCE & GENERALIZATION ABILITY"

4-2. What if the Bayesian Method Fails?

If it is not a good predictor (= no good relationship between the two), it indicates 2 things

- 1) numerical inaccuracies in the evaluation of the probabilities have caused the failure
- 2) alternative models (offered to Bayes) were a poor selection to the real world

(If we only used test error (w.o evidence), would have not been able to find the mismatch between model & data)

4-3. Back to the Demonstration : Comparing Different Regularizers

- Failure enables to progress with insight to new regularizers (find new prior that is more probable)
- better regularizers(=priors) can lead to a more correlated evidence & generalization error

###