

Machine Intelligence 1

3.3 Bayesian Inference and Neural Networks

Prof. Dr. Klaus Obermayer

Fachgebiet Neuronale Informationsverarbeitung (NI)

WS 2017/2018

3.3.1 Generative Models

Generative models

- observations: $\underline{\mathbf{z}}^{(\alpha)} = (\underline{\mathbf{x}}^{(\alpha)}, \underline{\mathbf{y}}_T^{(\alpha)})$ for $\alpha = 1, \dots, p$

$$p(\underline{\mathbf{z}}) = p(\underline{\mathbf{y}}_T | \underline{\mathbf{x}}) \cdot p(\underline{\mathbf{x}})$$

- most of our previous approaches:
 - ↪ construction of a parametrized class $y_{(\underline{\mathbf{x}}; \underline{\mathbf{w}})}$ of (deterministic) predictors
 - ↪ inference is based on ONE selected (optimal) predictor $y_{(\underline{\mathbf{x}}; \underline{\mathbf{w}}^*)}$

Generative models

- observations: $\underline{z}^{(\alpha)} = (\underline{x}^{(\alpha)}, \underline{y}_T^{(\alpha)})$ for $\alpha = 1, \dots, p$

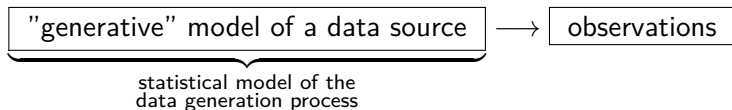
$$p(\underline{z}) = p(\underline{y}_T | \underline{x}) \cdot p(\underline{x})$$

- most of our previous approaches:

- ↪ construction of a parametrized class $y(\underline{x}; \underline{w})$ of (deterministic) predictors
 - ↪ inference is based on ONE selected (optimal) predictor $y(\underline{x}; \underline{w}^*)$

generative model approach:

- ↪ construction of a parametrized class $p(\underline{y} | \underline{x}; \underline{w})$ of (conditional) densities
- ↪ inference is based on good "generative models"



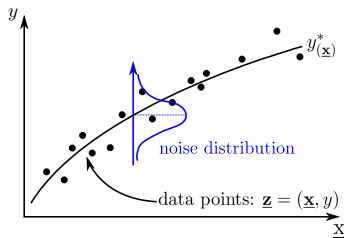
Comment

- The concept of generative models applies to supervised & unsupervised learning problems
- models $p(\underline{z}; \underline{w})$ for unconditional densities \leadsto unsupervised learning (e.g. ICA, mixture models)
- models $p(\underline{y} | \underline{x}; \underline{w})$ for conditional densities \leadsto supervised learning (e.g. “soft classification”)

Example I: Generative models for regression

Statistical of the data generation process:

$$y(\underline{x}) = \underbrace{y^*(\underline{x})}_{\text{deterministic relationship}} + \underbrace{\eta}_{\text{zero-mean noise}}$$

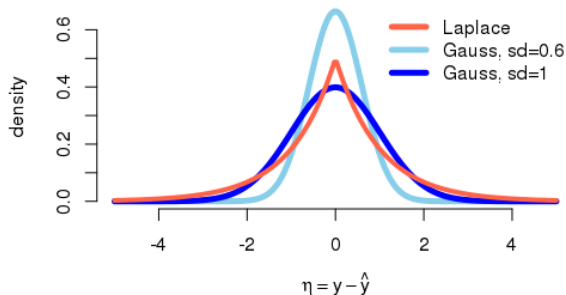


- Unknown deterministic relationship $y^*(\underline{x})$ approximated by parametrized function $\hat{y}(\underline{x}; \underline{w})$ (e.g. an ANN).
- Unknown noise process η approximated by parametrized distribution $\hat{p}(\eta; \sigma)$
- Here: additive noise.
 - other noise models possible (e.g. multiplicative noise)

Common noise models: Minkowski noise

- noise distribution is given as:

$$\hat{p}_{(y|\underline{x};\underline{w})} = \underbrace{\frac{d\beta^{\frac{1}{d}}}{2\Gamma(\frac{1}{d})}}_{\text{Gamma function}} \exp \left\{ -\beta |y - \hat{y}_{(\underline{x};\underline{w})}|^d \right\}$$



- $d = 1$: Laplace distribution, $d = 2$: Gaussian distribution

Example II: Classification for M classes C_k

Description of the data generation process

$$p(C_k|\underline{\mathbf{x}})$$

→ overlapping classes can induce 'label noise'

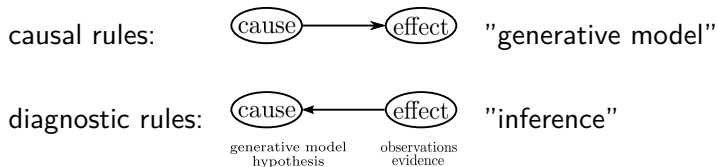
Model

$$\hat{p}(C_k|\underline{\mathbf{x}};\underline{\mathbf{w}}) = y_k(\underline{\mathbf{x}};\underline{\mathbf{w}})$$

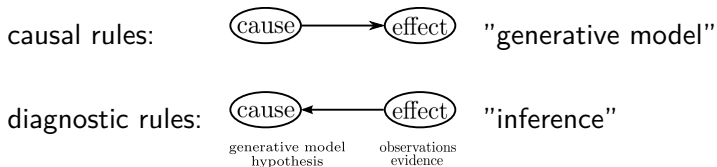
→ parametrized function (e.g. an ANN)

3.3.2 Bayesian Model Selection

Degrees of belief



Degrees of belief



Bayes rule

$$\underbrace{P_{(M|E)}}_{\text{posterior}} = \frac{\overbrace{P_{(E|M)}}^{\text{likelihood}} \overbrace{P_{(M)}}^{\text{prior}}}{\underbrace{P_{(E)}}_{\text{normalization constant ("evidence")}}}$$

Likelihood and prior

$$P(M_i|E) = \frac{P(E|M_i) P(M_i)}{P(E)}$$

Likelihood $P(E|M_i)$: probability of observing the evidence E ,
given that model M_i is true \Leftarrow **generative model**

Prior $P(M_i)$: degree of belief in M_i **before** E has been observed
initialization of prior beliefs \rightarrow maximum entropy methods

$$-\sum_i P_{(M_i)} \ln P_{(M_i)} \stackrel{!}{=} \max \quad (\text{least informative prior belief})$$

Constraints on the prior

$$\sum_i P_{(M_i)} = 1; \quad P_{(M_i)} \geq 0$$

uninformative prior: $P_{(M_i)} = \text{const.} \quad \rightsquigarrow \quad P_{(M_i|E)} \sim P_{(E|M_i)}$

Further constraints might be deduced from additional prior knowledge
e.g. about the value for the moments of $P(M_i)$. (see blackboard)

3.3.3 Bayesian Prediction

Bayesian committees

■ fundamental problem of prediction

observations E \longrightarrow degree of belief $P_{(e|E)}$
for a new event e

Bayesian committees

■ fundamental problem of prediction

observations E \longrightarrow degree of belief $P_{(e|E)}$
for a new event e



$$\begin{aligned}
 P_{(e|E)} &= \sum_i P_{(e, M_i|E)} && \text{marginalization} \\
 &= \sum_i P_{(e|M_i, E)} P_{(M_i|E)} && \text{def. of conditional probability} \\
 &\stackrel{!}{\approx} \sum_i P_{(e|M_i)} P_{(M_i|E)} && \text{conditional independence assumption}
 \end{aligned}$$

■ Bayesian committee: $P_{(e|E)} \approx \sum_i \underbrace{P_{(e|M_i)}}_{\text{likelihood}} \underbrace{P_{(M_i|E)}}_{\text{posterior}}$

Decision making: Minimizing expected loss

Cost of making a wrong prediction

$$C(\underbrace{e}_{\text{true}}, \underbrace{\hat{e}}_{\text{predicted}}) \quad \text{e.g.} \quad |e - \hat{e}|^d$$

Examples: 0-1 loss, squared error, absolute error, robust error criterion

- Decide for the value that minimizes the expected loss

$$\hat{e} = \operatorname{argmin}_{\tilde{e}} \int C_{(e, \tilde{e})} P_{(e|E)} de$$

- Decision for the most probable value only, if all errors are equally costly.

(see also Section 1.4.7)

3.3.4 Application: MLPs with Weight Decay

Recap: Bayes' theorem

$$\underbrace{P(M_i|E)}_{\text{posterior}} = \frac{\overbrace{P(E|M_i)}^{\text{likelihood}} \overbrace{P(M_i)}^{\text{prior}}}{\underbrace{P(E)}_{\text{normalization constant ("evidence")}}}$$

Construction of the data likelihood

training data: $\left\{ \left(\underline{\mathbf{x}}^{(\alpha)}, y_T^{(\alpha)} \right) \right\}, \alpha \in \{1, \dots, p\}$, *abbreviations:* $X = \left\{ \underline{\mathbf{x}}^{(\alpha)} \right\}, Y = \left\{ y_T^{(\alpha)} \right\}$

Data likelihood

ansatz:
$$P(y_T^{(\alpha)} | \underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}) \sim \exp \left(- \beta e_{(y_T^{(\alpha)}, \underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}})} \right)$$

Construction of the data likelihood

training data: $\left\{ \left(\underline{\mathbf{x}}^{(\alpha)}, y_T^{(\alpha)} \right) \right\}, \alpha \in \{1, \dots, p\}$, abbreviations: $X = \left\{ \underline{\mathbf{x}}^{(\alpha)} \right\}, Y = \left\{ y_T^{(\alpha)} \right\}$

Data likelihood

ansatz: $P(y_T^{(\alpha)} | \underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}) \sim \exp \left(-\beta e(y_T^{(\alpha)}, \underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}) \right)$

assumption: training data drawn i.i.d. from the joint distribution

$$\begin{aligned} P(Y | \underline{\mathbf{x}}; \underline{\mathbf{w}}) &\sim \prod_{\alpha} \exp \left(-\beta e(y_T^{(\alpha)}, \underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}) \right) \\ &\sim \exp \left(-\beta \sum_{\alpha} e(y_T^{(\alpha)}, \underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}) \right) \\ &\sim \exp \left(-\beta E_{(Y, X)}^T; \underline{\mathbf{w}} \right) \end{aligned}$$

Construction of the data likelihood

- Example: *additive Gaussian noise*

$$y_T^{(\alpha)} = \hat{y}_{(\underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}})} + \hat{\eta} \quad \text{with} \quad \hat{\eta} \sim \mathcal{N}(0, \sigma^2)$$

$$P(Y|\underline{\mathbf{x}}; \underline{\mathbf{w}}) = \frac{1}{(2\pi\sigma^2)^{\frac{p}{2}}} \exp \left\{ -\frac{1}{2\sigma^2} \underbrace{\sum_{\alpha=1}^p}_{\substack{\text{iid} \\ \text{assumption}}} \left(y_T^{(\alpha)} - \underbrace{\hat{y}_{(\underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}})}}_{\rightarrow \text{MLP}} \right)^2 \right\}$$

Construction of the data likelihood

- Example: *additive Gaussian noise*

$$y_T^{(\alpha)} = \hat{y}_{(\underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}})} + \hat{\eta} \quad \text{with} \quad \hat{\eta} \sim \mathcal{N}(0, \sigma^2)$$

$$\begin{aligned}
 P_{(Y|\underline{\mathbf{x}}; \underline{\mathbf{w}})} &= \frac{1}{(2\pi\sigma^2)^{\frac{p}{2}}} \exp \left\{ - \underbrace{\frac{1}{\sigma^2}}_{\beta} \overbrace{\sum_{\alpha=1}^p \frac{1}{2} \left(y_T^{(\alpha)} - \hat{y}_{(\underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}})} \right)^2}^{\text{training error } E^T: \text{quadratic error}} \right\} \\
 &\quad \underbrace{e(y_T^{(\alpha)}, \underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}})}_{\text{individual loss}} \\
 &= \frac{1}{Z} \prod_{\alpha=1}^p \exp \{ - \beta e(y_T^{(\alpha)}, \underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}) \}
 \end{aligned}$$

- maximizing the likelihood $P_{(Y|\underline{\mathbf{x}}; \underline{\mathbf{w}})} \sim$ minimizing the quadratic error E^T

Choice of the prior

- **Goal:** find the most “unprejudiced” distribution consistent with our prior knowledge (“constraints”)

Ansatz: the maximum entropy method

$$-\sum_{\underline{\mathbf{w}}} P_{(\underline{\mathbf{w}})} \ln P_{(\underline{\mathbf{w}})} \stackrel{!}{=} \max$$

$$\sum_{\underline{\mathbf{w}}} P_{(\underline{\mathbf{w}})} = 1 \quad (\text{normalization})$$

$$\sum_{\underline{\mathbf{w}}} E_{(\underline{\mathbf{w}})}^R P_{(\underline{\mathbf{w}})} = E_0 \quad (\text{prior knowledge: an example})$$

- examples: weight decay $E_{(\underline{\mathbf{w}})}^R = \sum_i w_i^2$ or Lasso $E_{(\underline{\mathbf{w}})}^R = \sum_i |w_i|$

Choice of the prior

Solution using Lagrange multipliers

$$-\sum_{\underline{\mathbf{w}}} P_{(\underline{\mathbf{w}})} \ln P_{(\underline{\mathbf{w}})} + \lambda \left(\sum_{\underline{\mathbf{w}}} P_{(\underline{\mathbf{w}})} - 1 \right) - \alpha \left(\sum_{\underline{\mathbf{w}}} E_{(\underline{\mathbf{w}})}^R P_{(\underline{\mathbf{w}})} - E_0 \right) \stackrel{!}{=} \max$$

Choice of the prior

Solution using Lagrange multipliers

$$\begin{aligned} - \sum_{\underline{\mathbf{w}}} P_{(\underline{\mathbf{w}})} \ln P_{(\underline{\mathbf{w}})} + \lambda \left(\sum_{\underline{\mathbf{w}}} P_{(\underline{\mathbf{w}})} - 1 \right) - \alpha \left(\sum_{\underline{\mathbf{w}}} E_{(\underline{\mathbf{w}})}^R P_{(\underline{\mathbf{w}})} - E_0 \right) &\stackrel{!}{=} \max \\ - \ln P_{(\underline{\mathbf{w}})} - 1 + \lambda - \alpha E_{(\underline{\mathbf{w}})}^R &= 0 \\ \ln P_{(\underline{\mathbf{w}})} &= \lambda - 1 - \alpha E_{(\underline{\mathbf{w}})}^R \\ P_{(\underline{\mathbf{w}})} &\sim \exp \left(- \alpha E_{(\underline{\mathbf{w}})}^R \right) \end{aligned}$$

- λ is found through normalization of prior probabilities
→ equivalent to choosing a normalization factor
- α can be calculated - in principle - from the corresponding constraint, however, it is often used as a hyperparameter

Comments

- Maximum entropy methods provide the “least informative” prior distribution $P(\underline{\mathbf{w}})$ for a given model architecture
- Prior knowledge, however, is already implicitly included by:
 - ↪ choice of parametrization (i.e. the architecture of the model $\hat{y}_{(\underline{\mathbf{x}};\underline{\mathbf{w}})})$
 - ↪ choice of noise model

Computing the posterior

■ Bayes rule:

$$\begin{aligned}
 P(\underline{\mathbf{w}}|Y,X) &\sim P(Y|X;\underline{\mathbf{w}})P(\underline{\mathbf{w}}) \\
 &\sim \exp\left\{-\frac{1}{2\sigma^2}E^T - \alpha E^R\right\} = \exp\left(-\frac{1}{2\sigma^2}R\right)
 \end{aligned}$$

■ where:

$\alpha' = 2\alpha\sigma^2$, the more data points, the less important the prior becomes

$$R = \underbrace{E^T}_{\sim \# \text{data}} + \underbrace{\alpha' E^R}_{\sim \# \text{parameters}}$$

Example: Additive Gaussian noise and weight decay

$$P(\underline{\mathbf{w}}|Y, X) \sim \exp\left(-\frac{1}{\sigma^2}R\right)$$

$$\text{with } R = \frac{1}{2} \sum_{\alpha=1}^p \left(y_T^{(\alpha)} - \hat{y}_{(\underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}})}\right)^2 + \frac{\alpha'}{2} \sum_{k=1}^d w_k^2$$

- Maximizing the posterior $P(\underline{\mathbf{w}}|Y, X)$
 \leadsto minimizing the regularized training error R .

Recap: Section 1.4.6

$$R_{[\underline{\mathbf{w}}]} = \underbrace{E_{[\underline{\mathbf{w}}]}^T}_{\text{training error}} + \underbrace{\lambda E_{[\underline{\mathbf{w}}]}^R}_{\text{regularization term}} \stackrel{!}{=} \min$$

E^R : penalizes certain models \leadsto "soft" restrictions on model space

λ : regularization parameter; trade-off between observations and prior knowledge

3.3.5 The "maximum a posteriori" Method

Prediction by Bayesian committee

$$P_{(y|\underline{\mathbf{x}};Y,X)} = \int P_{(y|\underline{\mathbf{x}};\underline{\mathbf{w}})} P_{(\underline{\mathbf{w}}|Y,X)} d\underline{\mathbf{w}}$$

(see Bishop Chapter 5.7)

Prediction by Bayesian committee

$$P_{(y|\underline{\mathbf{x}};Y,X)} = \int P_{(y|\underline{\mathbf{x}};\underline{\mathbf{w}})} P_{(\underline{\mathbf{w}}|Y,X)} d\underline{\mathbf{w}} = \int P_{(\underline{\mathbf{w}}|\{Y,y\},\{X,\underline{\mathbf{x}}\})} d\underline{\mathbf{w}}$$

$$P_{(y|\underline{\mathbf{x}};\underline{\mathbf{w}})} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{\sigma^2} e_{(y,\underline{\mathbf{x}};\underline{\mathbf{w}})}^T\right)$$

$$P_{(\underline{\mathbf{w}}|Y,X)} = \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left(-\frac{1}{\sigma^2} \sum_{\alpha=1}^p e_{(y_T^{(\alpha)},\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})}^T - \alpha' E_{[\underline{\mathbf{w}}]}^R\right)$$

(see Bishop Chapter 5.7)

Prediction by Bayesian committee

$$P_{(y|\underline{\mathbf{x}};Y,X)} = \int P_{(y|\underline{\mathbf{x}};\underline{\mathbf{w}})} P_{(\underline{\mathbf{w}}|Y,X)} d\underline{\mathbf{w}}$$

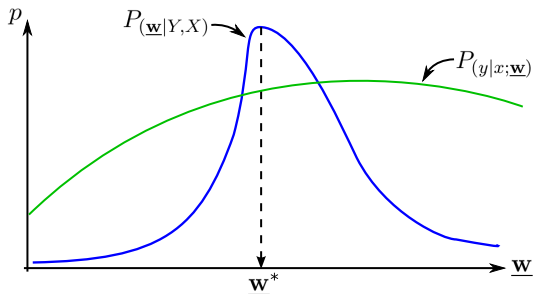
- There is no closed expression for the integral for many models.
- Numerical solutions, e.g. using MCMC methods.
- For some cases, the integral can be evaluated analytically.
 - regression with quadratic cost $e_{(y,\underline{\mathbf{x}};\underline{\mathbf{w}})}^T = \frac{1}{2}(y - \hat{y}_{(\underline{\mathbf{x}};\underline{\mathbf{w}})})^2$
 - linear functions $\hat{y}_{(\underline{\mathbf{x}};\underline{\mathbf{w}})} = \underline{\mathbf{w}}^\top \underline{\mathbf{x}}$
 - weight decay regularization $E_{[\underline{\mathbf{w}}]}^R = \frac{1}{2}\underline{\mathbf{w}}^\top \underline{\mathbf{w}}$
- Exact evaluation of the integral, but using approximations for the integrand.

(see Bishop Chapter 5.7)

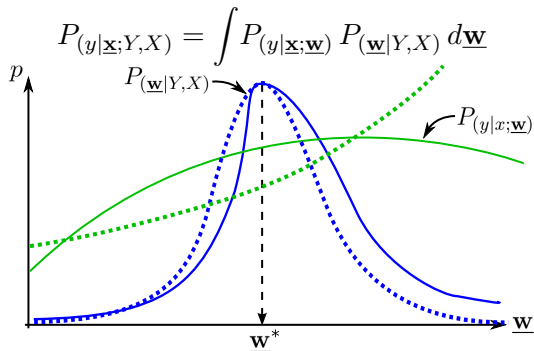
The maximum a posteriori approximation (MAP)

- **assumption:** Posterior has a **localized** maximum

$$P(y|\underline{x}; Y, X) = \int P(y|\underline{x}; \underline{w}) P(\underline{w}|Y, X) d\underline{w}$$



The maximum a posteriori approximation (MAP)



The maximum a posteriori approximation

$$P(y|\underline{\mathbf{x}}; Y, X) \sim \underbrace{\int \exp\left(-\frac{1}{2\sigma^2} e^T(\underline{\mathbf{w}}; y, x)\right)}_{\substack{\text{generative model} \\ \text{/ likelihood}}} \underbrace{\exp\left(-\frac{1}{2\sigma^2} R(Y, X; \underline{\mathbf{w}})\right)}_{\text{posterior}} d\underline{\mathbf{w}}$$

① Gauss-approximation of **posterior**

- Taylor expansion up to second order¹ around $\underline{\mathbf{w}}^*$

$$R(\underline{\mathbf{w}}, Y, X) = R(\underline{\mathbf{w}}^*, Y, X) + \frac{1}{2} \sum_{i,j} (\mathbf{w}_i - \mathbf{w}_i^*) \underbrace{\frac{\partial^2 R}{\partial \mathbf{w}_i \partial \mathbf{w}_j}}_{H_{ij}: \text{Hessian}} \bigg|_{\underline{\mathbf{w}}^*} (\mathbf{w}_j - \mathbf{w}_j^*)$$

② Linear approximation of the exponent of the individual **likelihood**

- Taylor expansion up to 1st order around $\underline{\mathbf{w}}^*$

$$e^T_{(y, \underline{\mathbf{x}}; \underline{\mathbf{w}})} = e^T_{(y, x; \underline{\mathbf{w}}^*)} + \sum_i \frac{\partial e^T}{\partial \mathbf{w}_i} \bigg|_{\underline{\mathbf{w}}^*} (\mathbf{w}_i - \mathbf{w}_i^*)$$

¹First order terms vanish, because $\underline{\mathbf{w}}^*$ is the location of the maximum.

The predictive distribution

- The MAP approximation yields an (approximate) closed form solution for the **predictive distribution**

$$P(y|\underline{\mathbf{x}}; Y, X) \sim \exp \left(-\beta e^T + \frac{\beta}{2} \left(\frac{\partial e^T}{\partial \underline{\mathbf{w}}} \right)^{\top} \underline{\mathbf{H}}^{-1} \frac{\partial e^T}{\partial \underline{\mathbf{w}}} \right) \bigg|_{\underline{\mathbf{w}}^*}$$

(calculation see supplementary material)

Example: MLP with weight decay

$$P_{(y|\underline{\mathbf{x}};Y,X)} \sim \exp \left(-\beta e^T + \frac{\beta}{2} \left(\frac{\partial e^T}{\partial \underline{\mathbf{w}}} \right)^\top \underline{\mathbf{H}}^{-1} \frac{\partial e^T}{\partial \underline{\mathbf{w}}} \right) \bigg|_{\underline{\mathbf{w}}^*}$$

■ example assumptions:

$$\beta = \frac{1}{\sigma^2} \quad e_{(\underline{\mathbf{x}};y;\underline{\mathbf{w}}^*)}^T = \frac{1}{2} (y - \hat{y}_{(\underline{\mathbf{x}};\underline{\mathbf{w}}^*)})^2 \quad \frac{\partial e^T}{\partial \underline{\mathbf{w}}} \bigg|_{\underline{\mathbf{w}}^*} = -(y - \hat{y}_{(\underline{\mathbf{x}};\underline{\mathbf{w}}^*)}) \underbrace{\frac{\partial \hat{y}}{\partial \underline{\mathbf{w}}} \bigg|_{\underline{\mathbf{w}}^*}}_{\underline{\mathbf{g}}}$$

■ approximate predictive distribution:

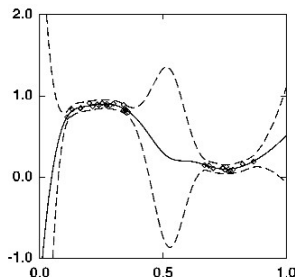
$$P_{(y|\underline{\mathbf{x}};Y,X)} \sim \exp \left\{ - \underbrace{\frac{1 - \underline{\mathbf{g}}^\top \underline{\mathbf{H}}^{-1} \underline{\mathbf{g}}}{\sigma^2}}_{1/\sigma_y^2} \frac{1}{2} (y - \hat{y}_{(\underline{\mathbf{x}};\underline{\mathbf{w}}^*)})^2 \right\}$$

Example: MLP with weight decay

$$P_{(y|\underline{\mathbf{x}}; Y, X)} \sim \exp \left\{ - \frac{1 - \underline{\mathbf{g}}^\top \underline{\mathbf{H}}^{-1} \underline{\mathbf{g}}}{\sigma^2} \frac{1}{2} (y - \hat{y}_{(\underline{\mathbf{x}}; \underline{\mathbf{w}}^*)})^2 \right\}$$

- predictive distribution is Gaussian with mean $\hat{y}_{(\underline{\mathbf{x}}; \underline{\mathbf{w}}^*)}$ and

$$\sigma_y^2 \stackrel{!}{=} \underbrace{1}_{\text{noise model}} - \underbrace{\frac{\underline{\mathbf{g}}^\top \underline{\mathbf{H}}^{-1} \underline{\mathbf{g}}}{\sigma^2}}_{\text{correction for parameter uncertainty}} \quad (\text{predictive variance})$$



Comments

- (1) $\underline{\mathbf{w}}^*$ is referred to as the "MAP-solution"

$$\underline{\mathbf{w}}^* = \underset{\underline{\mathbf{w}}}{\operatorname{argmin}} (E^T + \alpha' E^R)$$

Formal equivalence between MAP solution and regularized ERM:

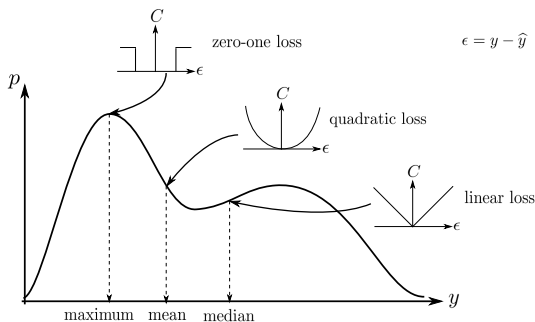
$$E^T \triangleq -\log \text{likelihood} \qquad E^R \triangleq -\log \text{prior}$$

- (2) For MLPs, both $\underline{\mathbf{g}}$ and $\underline{\mathbf{H}}^{-1}$ can be calculated efficiently (Bishop 2006)
- (3) $P(y|\underline{\mathbf{x}}; Y, X) \sim \exp(-\beta e^T)|_{\underline{\mathbf{w}}^*}$ is sometimes referred to as the *MAP solution* for the output distribution
- (4) The MAP solution accounts for two types of uncertainty
- uncertainty inherent in the generating process (σ^2)
 - precision of the estimated model ($1 - \underline{\mathbf{g}}^T \underline{\mathbf{H}}^{-1} \underline{\mathbf{g}}$)

Application: point prediction of attributes

- Find the prediction \hat{y} that minimizes the *expected loss*
 - for a given cost function $C_{(y,\tilde{y})}$
 - and given the probabilistic prediction $P_{(y|\underline{x};\underline{w})}$.

$$\hat{y}(\underline{x}) = \underset{\tilde{y}}{\operatorname{argmin}} \int dy C_{(y,\tilde{y})} P_{(y|\underline{x};\underline{w})}$$



- Gaussian distribution: maximum $\hat{=}$ mean $\hat{=}$ median