Machine Intelligence 1

3.3 Bayesian Inference and Neural Networks

Prof. Dr. Klaus Obermayer

Fachgebiet Neuronale Informationsverarbeitung (NI)

WS 2016/2017

3.3.1 Generative Models

Generative models

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

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

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

Generative models

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

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

- 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 model approach:

⤳ construction of a parametrized class $p_{(\underline{\mathbf{y}}|\underline{\mathbf{x}};\underline{\mathbf{w}})}$ of (conditional) densities

⤳ inference is based on good "generative models"

$$\underbrace{\boxed{\text{"generative" model of a data source}}}_{\substack{\text{statistical model of the} \\ \text{data generation process}}} \longrightarrow \boxed{\text{observations}}$$

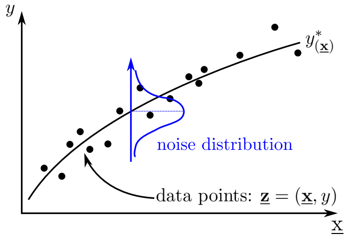
Comment

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

Example I: Generative models for regression

Statistical of the data generation process:

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

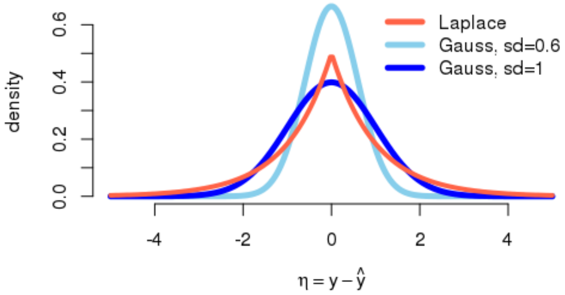


- Unknown deterministic relationship $y^*_{(\underline{\mathbf{x}})}$ approximated by parametrized function $\hat{y}_{(\underline{\mathbf{x}};\underline{\mathbf{w}})}$ (e.g. an ANN).
- Unknown noise process η approximated by parametrized distribution $\hat{p}(\eta; \sigma)$
- Here: additive noise.
 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

Common noise models: Minkowski noise

- noise distribution is given as:

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



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

Example II: Classification for M classes C_k

Description of the data generation process

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

$\rightsquigarrow$ overlapping classes can induce 'label noise'

Model

$$\widehat{p}_{(C_k|\underline{\mathbf{x}};\underline{\mathbf{w}})} = y_{k(\underline{\mathbf{x}};\underline{\mathbf{w}})}$$

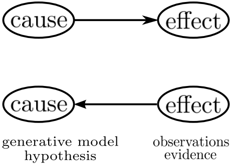
$\rightsquigarrow$ parametrized function (e.g. an ANN)

3.3.2 Bayesian Model Selection

Degrees of belief

causal rules: ''generative model''

diagnostic rules: ''inference''



Degrees of belief

causal rules: cause $\longrightarrow$ effect "generative model"

diagnostic rules: cause $\longleftarrow$ effect "inference"

generative model hypothesis | observations evidence

Bayes rule

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

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

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

Constraints on the prior

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

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

$$\text{observations } E \longrightarrow \begin{array}{c}\text{degree of belief } P_{(e|E)} \\ \text{for a new event } e\end{array}$$

Bayesian committees

- fundamental problem of prediction

$$\text{observations } E \longrightarrow \begin{array}{c}\text{degree of belief } P_{(e|E)} \\ \text{for a new event } e\end{array}$$

$$\begin{tikzcd} E \arrow[r] \arrow[rr, bend left] & M_i \arrow[r] & e \end{tikzcd} \quad \Rightarrow \quad \begin{tikzcd} E \arrow[r] & M_i \arrow[r] & e \end{tikzcd}$$

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

- 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

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

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

- Decide for the value that minimizes the expected loss

$$\widehat{e} \quad = \quad \underset{\tilde{e}}{\operatorname{argmin}} \int C_{(e,\tilde{e})} P_{(e|E)} \, de$$

- 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

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

Construction of the data likelihood

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

Data likelihood

ansatz:

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

Construction of the data likelihood

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

Data likelihood

ansatz:

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

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

$$\begin{aligned} P_{(Y | \underline{\mathbf{x}}; \underline{\mathbf{w}})} &\sim \prod_{\alpha} \exp\left(-\beta \, e_{\left(y_T^{(\alpha)}, \underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}\right)} \right) \\ &\sim \exp\left(-\beta \sum_{\alpha} e_{\left(y_T^{(\alpha)}, \underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}}\right)} \right) \\ &\sim \exp\left(-\beta \, E^T_{(Y, X; \underline{\mathbf{w}})} \right) \end{aligned}$$

Construction of the data likelihood

- Example: *additive Gaussian noise*

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

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

Construction of the data likelihood

- Example: *additive Gaussian noise*

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

$$P_{(Y|\underline{\mathbf{x}};\underline{\mathbf{w}})} = \frac{1}{(2\pi\sigma^2)^{\frac{p}{2}}} \exp\Bigg\{ -\underbrace{\frac{1}{\sigma^2}}_{\beta} \overbrace{\sum_{\alpha=1}^{p} \underbrace{\tfrac{1}{2}\Big(y_T^{(\alpha)} - \widehat{y}_{(\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})}\Big)^2}_{\substack{\text{individual loss} \\ e(y_T^{(\alpha)},\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})}}}^{\text{training error } E^T\text{: quadratic error}} \Bigg\}$$

$$= \frac{1}{Z} \prod_{\alpha=1}^{p} \exp\left\{ -\beta\, e(y_T^{(\alpha)}, \underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}})\right\}$$

- 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

$$
\begin{aligned}
-\sum_{\underline{\mathbf{w}}} P_{(\underline{\mathbf{w}})} \ln P_{(\underline{\mathbf{w}})} &\overset{!}{=} \max \\
\sum_{\underline{\mathbf{w}}} P_{(\underline{\mathbf{w}})} &= 1 \qquad \text{(normalization)} \\
\sum_{\underline{\mathbf{w}}} E^R_{(\underline{\mathbf{w}})} P_{(\underline{\mathbf{w}})} &= E_0 \qquad \text{(prior knowledge: an example)}
\end{aligned}
$$

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

Choice of the prior

Solution using Lagrange multipliers

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

Choice of the prior

Solution using Lagrange multipliers

$$-\sum_{\underline{\mathbf{w}}} P_{(\underline{\mathbf{w}})} \ln P_{(\underline{\mathbf{w}})} + \lambda \Bigg(\sum_{\underline{\mathbf{w}}} P_{(\underline{\mathbf{w}})} - 1 \Bigg) - \alpha \Bigg(\sum_{\underline{\mathbf{w}}} E^R_{(\underline{\mathbf{w}})} P_{(\underline{\mathbf{w}})} - E_0 \Bigg) \stackrel{!}{=} \max$$

$$\begin{aligned} -\ln P_{(\underline{\mathbf{w}})} - 1 + \lambda - \alpha E^R_{(\underline{\mathbf{w}})} &= 0 \\ \ln P_{(\underline{\mathbf{w}})} &= \lambda - 1 - \alpha E^R_{(\underline{\mathbf{w}})} \\ P_{(\underline{\mathbf{w}})} &\sim \exp\Big(-\alpha E^R_{(\underline{\mathbf{w}})} \Big) \end{aligned}$$

- λ is found through normalization of prior probabilities
  $\rightarrow$ 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:
  - $\rightsquigarrow$ choice of parametrization (i.e. the architecture of the model $\hat{y}_{(\underline{\mathbf{x}};\underline{\mathbf{w}})}$)
  - $\rightsquigarrow$ choice of noise model

Computing the posterior

- Bayes rule:

$$P_{(\underline{\mathbf{w}}|Y,X)} \sim P_{(Y|X;\underline{\mathbf{w}})} P_{(\underline{\mathbf{w}})}$$

$$\sim \exp\left\{ -\frac{1}{2\sigma^2} E^T - \alpha E^R \right\} \quad = \quad \exp\left(-\frac{1}{2\sigma^2} R\right)$$

- 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

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

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

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

Recap: Section 1.4.6

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

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

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

$$\begin{aligned} P_{(y|\underline{\mathbf{x}};\underline{\mathbf{w}})} &= \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{1}{\sigma^2} e^T_{(y,\underline{\mathbf{x}};\underline{\mathbf{w}})}\Big) \\ P_{(\underline{\mathbf{w}}|Y,X)} &= \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\Big(-\frac{1}{\sigma^2} \sum_{\alpha=1}^{p} e^T_{(y_T^{(\alpha)},\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} - \alpha' E^R_{[\underline{\mathbf{w}}]}\Big) \end{aligned}$$

(see Bishop Chapter 5.7)

Prediction by Bayesian committee

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

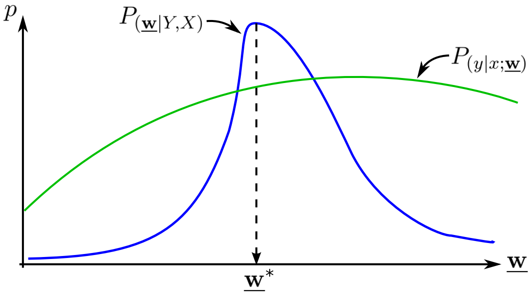
- There is no closed expression for the integral for many models.
- Numerical solutions, e.g. using MCMC methods.
- For some cases, the integral can be evaluated analytically.
  - regression with quadratic cost $e^T_{(y,\underline{\mathbf{x}};\underline{\mathbf{w}})} = \frac{1}{2}(y - \hat{y}_{(\underline{\mathbf{x}};\underline{\mathbf{w}})})^2$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

(see Bishop Chapter 5.7)

The maximum a posteriori approximation (MAP)

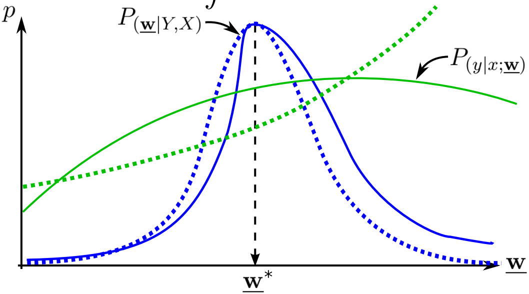
- **assumption:** Posterior has a localized maximum

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



The maximum a posteriori approximation (MAP)

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



The maximum a posteriori approximation

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

1. Gauss-approximation of posterior
   - Taylor expansion up to second order[1] around $\underline{\mathbf{w}}^*$

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

2. Linear approximation of the exponent of the individual likelihood
   - Taylor expansion up to 1st order around $\underline{\mathbf{w}}^*$

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

[1] First 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:

$$\beta = \tfrac{1}{\sigma^2} \qquad e^T_{(\underline{\mathbf{x}},y;\underline{\mathbf{w}}^*)} = \tfrac{1}{2}\left(y - \hat{y}_{(\underline{\mathbf{x}};\underline{\mathbf{w}}^*)}\right)^2 \qquad \frac{\partial e^T}{\partial \underline{\mathbf{w}}}\bigg|_{\underline{\mathbf{w}}^*} = -\left(y - \widehat{y}_{(\underline{\mathbf{x}};\underline{\mathbf{w}})}\right) \underbrace{\tfrac{\partial \widehat{y}}{\partial \underline{\mathbf{w}}}\Big|_{\underline{\mathbf{w}}^*}}_{\underline{\mathbf{g}}}$$

- approximate predictive distribution:

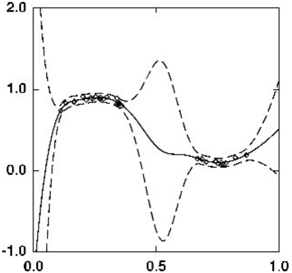
$$P_{(y|\underline{\mathbf{x}};Y,X)} \sim \exp\left\{ -\underbrace{\frac{1 - \underline{\mathbf{g}}^\top \underline{\mathbf{H}}^{-1} \underline{\mathbf{g}}}{\sigma^2}}_{1/\sigma_y^2} \tfrac{1}{2}\left(y - \widehat{y}_{(\underline{\mathbf{x}};\underline{\mathbf{w}}^*)}\right)^2 \right\}$$

Example: MLP with weight decay

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

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

$$\sigma_y^2 \overset{!}{=} \frac{\sigma^2}{\underbrace{1}_{\text{noise model}} - \underbrace{\underline{\mathbf{g}}^\top \underline{\mathbf{H}}^{-1}\underline{\mathbf{g}}\Big|_{\underline{\mathbf{w}}^*}}_{\text{correction for parameter uncertainty}}} \qquad \text{(predictive variance)}$$



Comments

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

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

Formal equivalence between MAP solution and regularized ERM:

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

(2) For MLPs, both $\underline{\mathbf{g}}$ and $\underline{\mathbf{H}}^{-1}$ can be calculated efficiently (Bishop 2006)

(3) $P_{(y|\underline{\mathbf{x}};Y,X)} \sim \exp(-\beta e^T)\big|_{\underline{\mathbf{w}}^*}$ is sometimes referred to as the *MAP solution* for the output distribution

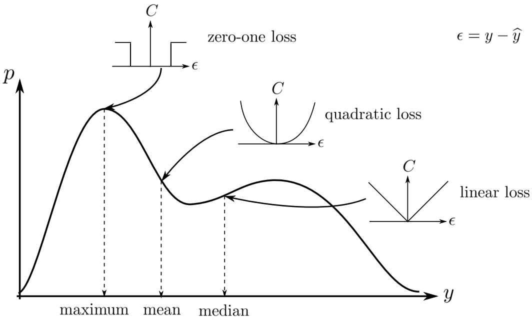
(4) The MAP solution accounts for two types of uncertainty

- 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 $P_{(y|\underline{\mathbf{x}};\underline{\mathbf{w}})}$.

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



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