

MAP 534

Introduction to machine learning

Linear models for classification II

Alain Durmus

Probabilistic discriminative models for classification

Logistic regression

Bayesian logistic regression

Introduction to classification: setting

- We consider a supervised setting.
- Decide how to encode inputs and outputs: this defines the input space X , and the output space Y .
- Here we consider specifically the classification problem: Y is a finite set,

$Y = \{1, \dots, K\}$, in most of this lecture even $Y = \{0, 1\}$.

- In this lecture, we apply the three machine learning paradigms to address:

$$y_{\text{pred}} = \mathcal{C}(x_{\text{new}}) ,$$

and aim to quantify if possible the uncertainties of our predictions.

- Here \mathcal{C} is called a classifier.
- Recall that the three paradigms are:
 - deterministic discriminative learning ✓;
 - probabilistic generative learning ✓;
 - probabilistic discriminative learning \implies today;

Generative vs discriminative models

- Recall the generative setting: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ are i.i.d observations of random variables (X, Y) :

$$X|Y \sim p_w(\cdot|Y) , Y \sim p_Y .$$

- We consider here a family of conditional densities

$$\{(x, y) \mapsto p_w(x|y) : w \in \Theta\} .$$

- From an estimator \hat{w} , we end up using Bayes formula with the predictive distribution

$$p_{\hat{w}}(y|x_{\text{new}}) = \frac{p_{\hat{w}}(x|y)p_Y(y)}{\sum_{y' \in Y} p_{\hat{w}}(x|y')p_Y(y')} \propto p_{\hat{w}}(x|y)p_Y(y) .$$

- Idea of discriminative models: directly consider a family of conditional distributions

$$\{(y, x) \mapsto p_w(y|x) : w \in \Theta\} .$$

Discriminative models

- Idea of discriminative models: directly consider a family of conditional distributions

$$\{(y, x) \mapsto p_w(y|x) : w \in \Theta\} .$$

- Choosing a prior $p_X(x)$, we get the likelihood


$$L(w; \mathcal{D}) = \prod_{i=1}^N \{p_w(y_i|x_i)p_X(x_i)\} .$$

- In most cases, we do not infer/learn p_X .
- Then, w is estimated by maximum likelihood which is equivalent in that case to minimize:

$$E(w) = - \sum_{i=1}^N \log p_w(y_i|x_i) .$$

- We end up with estimating the predictive probabilities $p_{\hat{w}}(y|x_{\text{new}})$ and the prediction:

$$y_{\text{pred}} = \operatorname{argmax}_y p_{\hat{w}}(y|x_{\text{new}}) .$$

-  how to choose the family $\{(y, x) \mapsto p_w(y|x) : w \in \Theta\}$?
- We can first take inspiration from the form we obtain in generative models.

Introduction and motivation: logistic and softmax functions

- Recall the generative setting: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ are i.i.d observations of random variables (X, Y) :

$$X|Y \sim p_w(\cdot|Y), Y \sim p_Y.$$

- We consider here a family of conditional densities

$$\{(x, y) \mapsto p_w(x|y) : w \in \Theta\}.$$

- From an estimator \hat{w} , we end up using Bayes formula with the predictive distribution

$$p_{\hat{w}}(y|x_{\text{new}}) = \frac{p_{\hat{w}}(x|y)p_Y(y)}{\sum_{y' \in \mathcal{Y}} p_{\hat{w}}(x|y')p_Y(y')} \propto p_{\hat{w}}(x|y)p_Y(y).$$

- In the case $\mathcal{Y} = \{0, 1\}$, this can be written on the form:

$$p_{\hat{w}}(1|x_{\text{new}}) = \sigma(a(x_{\text{new}}, \hat{w})), \quad \sigma(t) = \frac{e^t}{1 + e^t}$$

$$a(x_{\text{new}}, w) = \log \left(\frac{p_w(x_{\text{new}}|1)p_{\text{prior}}(1)}{p_w(x_{\text{new}}|0)p_{\text{prior}}(0)} \right)$$

- And similarly for $y = 0$...
- The quantity $a(x, w)$ is called an activation.
- The function σ is called the logistic sigmoid function.

The logistic sigmoid function

- $\sigma(t) = \frac{e^t}{1+e^t}$ very important function in ML.
- Some properties:

$$\sigma(-t) = 1 - \sigma(t) , \quad \text{inverse } \sigma \mapsto \log(\sigma/(1 - \sigma)) .$$

Proof.



Introduction and motivation: logistic and softmax functions

- Recall the generative setting: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ are i.i.d observations of random variables (X, Y) :

$$X|Y \sim p_w(\cdot|Y), Y \sim p_Y.$$

- We consider here a family of conditional densities $\{(x, y) \mapsto p_w(x|y) : w \in \Theta\}$.
- In the case $Y = \{0, 1\}$, the predictive probability can be written of the form:

$$p_w(1|x_{\text{new}}) = \sigma(a(x_{\text{new}}, w)), \quad \sigma(t) = \frac{e^t}{1 + e^t}$$
$$a(x_{\text{new}}, w) = \log \left(\frac{p_w(x_{\text{new}}|1)p_{\text{prior}}(1)}{p_w(x_{\text{new}}|0)p_{\text{prior}}(0)} \right)$$

- The activation $a(x_{\text{new}}, \hat{w}) = \log \text{odds} / \log \text{ of ratio of posterior probabilities}$
- The expression for $p_{\hat{w}}$ seems to be overly complicated given the first simple expression given previously...
- However it gives ideas of many discriminative models: choose a !
- Idea of logistic regression: take $a(x_{\text{new}}, w) = \phi(x_{\text{new}})^T w$, linear with respect to the parameter.

Multiclass discriminative probabilistic problem

- This generalizes to the multiclass scenario $K > 2$.
- In that situation: we end up with predictive probabilities for the class k of the form:

$$p_W(k|x_{\text{new}}) = \frac{\exp(a(x_{\text{new}}, w_k))}{\sum_{j=1}^K \exp(a(x_{\text{new}}, w_j))},$$

where $W = \{w_j\}_{j=1}^K$ is the parameter to infer.

- The function $\sigma : \mathbb{R}^K \rightarrow \text{Simplex}_K = \{(\varpi_1, \dots, \varpi_K) \in [0, 1]^K : \sum_{k=1}^K \varpi_k = 1\}$:

$$\sigma(a_1, \dots, a_K) = \left(\frac{e^{a_1}}{\sum_{j=1}^K e^{a_j}}, \dots, \frac{e^{a_K}}{\sum_{j=1}^K e^{a_j}} \right)^T$$

is called the normalized exponential or softmax function.

- With this notation, $p_W(k|x_{\text{new}}) = \sigma(a_1, \dots, a_K)_k$, with $a_k = a(x_{\text{new}}, w_k)$.
- This remark suggests also to directly choose the function $a : X \times \mathbb{R}^d \rightarrow \mathbb{R}$.

Probabilistic discriminative models for classification

Logistic regression

Bayesian logistic regression

- We consider here fixed basis functions $\phi = \{\phi_j\}_{j=1}^d$, $\phi_j : X \rightarrow \mathbb{R}$.
- This helps in the process of modeling the predictive distribution from activations.
- Without loss of generality, we then consider that the feature $x_i \in \mathbb{R}^d$ changing $x_i \leftarrow \{\phi_1(x_i), \dots, \phi_d(x_i)\}$.

Logistic regression: likelihood

- We first consider the 0 – 1 prediction task, $Y = \{0, 1\}$.
- We consider here the family of conditional/predictive distributions:
 $w \in \mathbb{R}^d$,

$$p_w(y_{\text{pred}}|x_{\text{new}}) = \sigma(w^T x_{\text{new}})^{y_{\text{pred}}} (1 - \sigma(w^T x_{\text{new}}))^{1-y_{\text{pred}}} , \quad \sigma(t) = \frac{e^t}{1 + e^t} .$$

- This corresponds to the statistical model: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ are i.i.d observations from

$$Y|X \sim \text{Ber}(\sigma(w^T X)) .$$

- The likelihood is then given by


$$L(w; \mathcal{D}) = ?? .$$

- Note that the distribution on X is here not relevant.

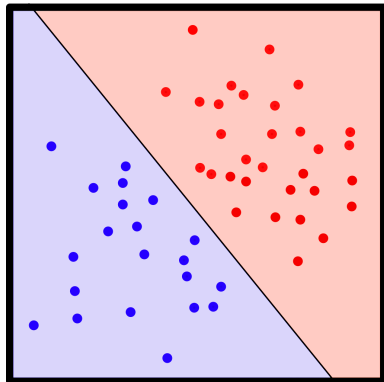
MLE for the logistic regression

- Maximizing L is equivalent to minimizing the negative log-likelihood which gives rise to the error function:

$$E(w) = - \sum_{i=1}^N \{y_i \log(\sigma_i(w)) + (1 - y_i) \log(1 - \sigma_i(w))\} , \quad \sigma_i = \sigma(w^T x_i) .$$

- E is called the cross-entropy error function.
- The gradient of E is given by 
- The problem of minimizing E does not admit explicit solutions.
- We must use optimization algorithms such as GD or SGD.

Overfitting phenomenon for the logistic regression



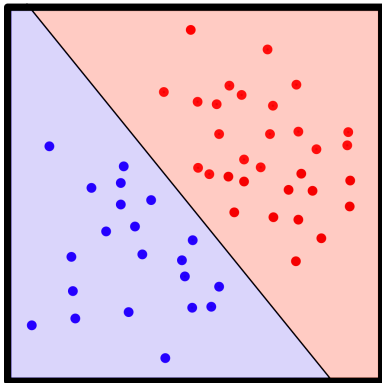
- In the case the data is linearly separable: i.e., there exists $\bar{\mathbf{w}} \in \mathbb{R}^d$ such that

$$\{\mathbf{x}_i : y_i = 1\} \subset H_- = \{\mathbf{x} : \bar{\mathbf{w}}^T \mathbf{x} < 0\},$$
$$\{\mathbf{x}_i : y_i = 0\} \subset H_+ = \{\mathbf{x} : \bar{\mathbf{w}}^T \mathbf{x} > 0\}.$$

then MLE for logistic regression exhibits severe overfitting.

- MLE tends to provide parameters $\hat{\mathbf{w}} = \alpha \bar{\mathbf{w}}$, $\alpha \in \mathbb{R}$, with $|\alpha|$ very large.
- This results in predictive $p_{\hat{\mathbf{w}}}(1|\mathbf{x}_{\text{new}})$ which tends to the non-smooth Heaviside function: $\mathbb{1}_{\{\bar{\mathbf{w}}^T \mathbf{x} < 0\}}$.
- Problem remains even as N is large.

Overfitting phenomenon for the logistic regression



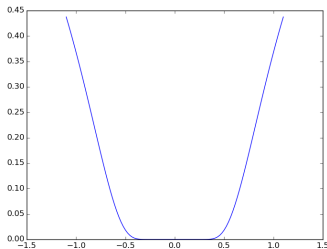
- In the case the data is linearly separable: i.e., there exists $\bar{w} \in \mathbb{R}^d$ such that

$$\{x_i : y_i = 1\} \subset H_- = \{x : \bar{w}^T x < 0\},$$
$$\{x_i : y_i = 0\} \subset H_+ = \{x : \bar{w}^T x > 0\}.$$

then MLE for logistic regression exhibits severe overfitting.

- Reason: in that situation, the MLE is not unique and there exists an infinite numbers of solutions.
- Standard MLE procedure does not provide a way to favor one solution over another.
- Solution: take a prior on w and consider the MAP instead of the MLE.

Vanishing gradient for the logistic regression



- GD for logistic regression can be slow since E exhibits flat regions;
- this implies vanishing gradients...
- To better understand this phenomenon and accelerate the convergence, we make use of the Hessian of E :


$$\nabla_w^2 E(w) = \left(\frac{\partial^2 E}{\partial w_i \partial w_j}(w) \right)_{i,j=1}^d$$

Convexity of the logistic regression and vanishing gradient phenomenon

- From the expression for $\nabla^2 E$, we can show that E is convex

Proof.



- However, $\min \text{Spec} \nabla^2 E(w) \approx 0$ and also $\nabla E(w) \approx 0$ near minimums.
This mathematically formalizes the previous statements.
- Therefore, while E is convex, the convergence will be very slow.
- To get faster convergence, we need to take into account this observation.
- Example .

- Newton algorithm aims to deal with the problem of vanishing gradient.
- It acts as an adaptive pre-conditioned gradient descent scheme.
- Pre-conditioned gradient descent scheme?

Iterated reweighted least square/ Newton algorithm

- Newton algorithm aims to deal with the problem of vanishing gradient.
- It acts as an adaptive pre-conditioned gradient descent scheme.
- It defines the recursion:

$$w_{k+1} = w_k - \{\nabla^2 E(w_k)\}^{-1} \nabla E(w_k) .$$

- Example  .

Multiclass logistic regression

- We aim to generalize the logistic regression for $Y = \{1, \dots, K\}$, $K > 2$.
- Recall that we start with the statistical model:

$$Y \sim \text{Ber}(\sigma(w, X)) ,$$

where σ is an activation function.

- This generalizes easily as

$$Y \sim \text{Multi}(\sigma_1(W, X), \dots, \sigma_K(W, X)) , \quad W = [w_1 \cdots w_K] \in \mathbb{R}^{d \times K} ,$$

where

$$\sigma : \mathbb{R}^{d \times K} \times X \rightarrow \text{Simplex}_K = \{(\varpi_1, \dots, \varpi_K) \in [0, 1]^K : \sum_{k=1}^K \varpi_k = 1\} .$$

- This comes from our previous discussion on generative probabilistic models.

Multiclass logistic regression

- We consider the multiclass setting using logistic regression.
- The logistic model generalizes easily as

$$Y \sim \text{Multi}(\sigma_1(W, X), \dots, \sigma_K(W, X)) , \quad W = [w_1 \cdots w_K] \in \mathbb{R}^{d \times K} ,$$

where

$$\sigma : \mathbb{R}^{d \times K} \times X \rightarrow \text{Simplex}_K = \{(\varpi_1, \dots, \varpi_K) \in [0, 1]^K : \sum_{k=1}^K \varpi_k = 1\} ,$$

$$\sigma_k(W, X) = \text{softmax}(X^T w_1, \dots, X^T w_K)_k = \frac{\exp(X^T w_k)}{\sum_{j'=1}^K \exp(X^T w_{j'})} .$$

- This comes from our previous discussion on generative probabilistic models again...
- The likelihood is then given by (we skip as usual the prior on X)

$$L(W; \mathcal{D}) = \prod_{i=1}^N \mathbb{P}(Y_i = y_i | X = x_i) = \prod_{i=1}^N \prod_{k=1}^K \{\sigma_{k,i}\}^{y_i=k} ,$$

setting

$$\sigma_{k,i} = \sigma_k(W, x_i) .$$

Multiclass logistic regression

- The likelihood is then given by (we skip as usual the prior on X)

$$L(\mathbf{W}; \mathcal{D}) = \prod_{i=1}^N \mathbb{P}(Y_i = y_i | X = x_i) = \prod_{i=1}^N \prod_{k=1}^K \{\sigma_{k,i}\}^{y_i=k},$$

setting

$$\sigma_{k,i} = \sigma_k(\mathbf{W}, x_i).$$

- Using the 1-of- K coding for the labels, $t_{k,i} = \mathbb{1}\{y_i = k\}$ and taking the logarithm, we get

$$E(\mathbf{W}) = -\log L(\mathbf{W}; \mathcal{D}) = -\sum_{i=1}^N \sum_{k=1}^K t_{k,i} \log(\sigma_{k,i}).$$

-  $\nabla_{\mathbf{W}} E$ .

-  $\nabla_{\mathbf{W}}^2 E$ .

Probit regression

- We focus on a $\{0, 1\}$ classification task and consider a discriminative probabilistic model:

$$Y \sim \text{Ber}(\sigma(w^T X)) ,$$

where $\sigma : \mathbb{R} \rightarrow [0, 1]$ is an activation function.

- For the logistic regression, we simply take the logistic sigmoid function.
- However, we can take any cumulative distribution function on \mathbb{R} , F , for $\sigma \dots$
- The probit regression just consists in taking $\sigma = \Phi$ where Φ is the cumulative distribution function associated with $N(0, 1)$:

$$\Phi(t) = \int_{-\infty}^t e^{-u^2/2} du / (2\pi)^{1/2} .$$

- Exercise: write the likelihood associated with the probit regression model for $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$.
- Then w is inferred as the logistic regression by maximum likelihood estimation.
- Again the maximum of the likelihood does not admit a close form and gradient descent, SGD or Newton algorithm has to be used...

Probabilistic discriminative models for classification

Logistic regression

Bayesian logistic regression

Posterior distribution

- Recall that the likelihood associated with the logistic regression model for $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ is?
- Taking a prior for w , p_{prior} , we get using the Bayes formula with the posterior distribution

$$p(w|\mathcal{D}) \propto L(w; \mathcal{D}) \times p_{\text{prior}}(w) .$$

- Following the same discussion as the Bayesian linear regression, Bayesian classification is based on the posterior predictive distribution defined by

$$p(y_{\text{pred}}|x_{\text{new}}) = \int ?? .$$

- This integral is not tractable! because the posterior distribution cannot be sampled from.
- In the sequel, we present an approximation method of the posterior.

Laplace approximation method

- Here we consider a general target posterior distribution on \mathbb{R}^d

$$p(w|\mathcal{D}) = \lambda(w)/Z, \quad \lambda(w) = L(w; \mathcal{D})p_{\text{prior}}(w), \quad Z = \int \lambda(w)dw.$$

- Here the density is known up to a multiplicative constant: λ is known but not Z .
- The basic idea of Laplace approximation is to find μ and $\Sigma \succeq 0$ such that

$$p(\cdot|\mathcal{D}) \approx \mathcal{N}(\mu, \Sigma).$$

- Theoretical justification: Bernstein von Mises theorem shows that an appropriate scaled version of the posterior converges to a Gaussian distribution.



- which μ, Σ ?
- We consider for μ the MAP estimator:

$$w_{\text{MAP}} = \operatorname{argmax} p(\cdot|\mathcal{D}) = \operatorname{argmax} \lambda.$$

- Basic idea: $p(\cdot|\mathcal{D})$ concentrates around w_{MAP} .
- Remains the choice of Σ .

Laplace approximation method

- Here we consider a general target posterior distribution on \mathbb{R}^d

$$p(w|\mathcal{D}) = e^{-E(w)}/Z, \quad e^{-E(w)} = L(w; \mathcal{D})p_{\text{prior}}(w), \quad Z = \int \lambda(w)dw.$$

- The basic idea of Laplace approximation is to find μ and $\Sigma \succeq$ such that

$$p(\cdot|\mathcal{D}) \approx N(\mu, \Sigma).$$

- We consider for μ the MAP estimator:

$$w_{\text{MAP}} = \operatorname{argmax} p(\cdot|\mathcal{D}) = \operatorname{argmin} E.$$

- For Σ , consider first $d = 1$ and making a second Taylor expansion at w_{MAP} , we get

$$\begin{aligned} \log p(w|\mathcal{D}) &\approx \log p(w_{\text{MAP}}|\mathcal{D}) + [\partial_w \log p(w_{\text{MAP}}|\mathcal{D})](w - w_{\text{MAP}}) \\ &\quad + (1/2)[\partial_w^2 \log p(w_{\text{MAP}}|\mathcal{D})](w - w_{\text{MAP}})^2 \\ &\approx \log p(w_{\text{MAP}}|\mathcal{D}) - (1/2) E''(w_{\text{MAP}})(w - w_{\text{MAP}})^2. \end{aligned}$$

- This suggests to take

$$\Sigma = 1/E''(w_{\text{MAP}}), \text{ if } E''(w_{\text{MAP}}) \neq 0.$$

Laplace approximation method

- Here we consider a general target posterior distribution on \mathbb{R}^d

$$p(w|\mathcal{D}) = e^{-E(w)}/Z, \quad e^{-E(w)} = L(w; \mathcal{D})p_{\text{prior}}(w), \quad Z = \int e^{-E(w)} dw.$$

- The basic idea of Laplace approximation is to find μ and $\Sigma \succeq$ such that

$$p(\cdot|\mathcal{D}) \approx N(\mu, \Sigma).$$

- We consider for μ the MAP estimator:

$$w_{\text{MAP}} = \operatorname{argmax} p(\cdot|\mathcal{D}) = \operatorname{argmin} E.$$

- The previous derivation generalizes for $d > 1$ and we obtain

$$\Sigma = [\nabla^2 E(w_{\text{MAP}})]^{-1}.$$

- Note that we do not need Z to compute w_{MAP} and Σ since they can only be related to E .

Application to the Bayesian Logistic regression

- Recall that the posterior has a form

$$-\log p(w|\mathcal{D}) = E + \text{Cst} ,$$

$$E = \frac{1}{2}(w - m_0)\mathbf{S}_0^{-1}(w - m_0) - \sum_{i=1}^N \{y_i \log(\sigma_i) + (1 - y_i) \log(1 - \sigma_i)\}$$

where

$$\sigma_i = ??$$

- Assume that we have estimated the MAP, w_{MAP} , the Hessian for E is

$$\nabla^2 E(w_{\text{MAP}}) = ??$$

- Finally the Laplace approximation is

$$p_{\text{laplace}}(w|\mathcal{D}) = \mathbf{N}(w_{\text{MAP}}, [\nabla^2 E(w_{\text{MAP}})]^{-1})$$

- The Laplace approximation is

$$p_{\text{laplace}}(w|\mathcal{D}) = \mathcal{N}(w_{\text{MAP}}, [\nabla^2 E(w_{\text{MAP}})]^{-1})$$

- The predictive distribution is itself approximated by

$$\begin{aligned} p_{\text{pred}}(y|x_{\text{new}}) &= \int p_w(y|x) p(w|\mathcal{D}) dw \approx \int p_w(y|x) p_{\text{laplace}}(w|\mathcal{D}) dw \\ &= \mathbb{E} [\sigma(W^T x_{\text{new}})] , \quad W \sim \mathcal{N}(w_{\text{MAP}}, [\nabla^2 E(w_{\text{MAP}})]^{-1}) \\ &= \mathbb{E} [\sigma(G(x_{\text{new}}))] , \quad G(x_{\text{new}}) \sim ?? \text{🖊} . \end{aligned}$$

- $\mathbb{E} [\sigma(G(x_{\text{new}}))]$ has to be approximated, cannot be evaluated analytically.

- Recall that the posterior has a form

$$\begin{aligned} -\log p(w|\mathcal{D}) &= E + \text{Cst} , \\ E &= \dots , \end{aligned}$$

- Assume that we have estimated the MAP, w_{MAP} , the Hessian for E is

$$\nabla^2 E(w_{\text{MAP}}) = ??$$

- Finally the Laplace approximation is

$$p_{\text{laplace}}(w|\mathcal{D}) = \text{N}(w_{\text{MAP}}, [\nabla^2 E(w_{\text{MAP}})]^{-1}) .$$

- The Laplace approximation is

$$p_{\text{laplace}}(w|\mathcal{D}) = \mathcal{N}(w_{\text{MAP}}, [\nabla^2 E(w_{\text{MAP}})]^{-1})$$

- The predictive distribution is itself approximated by

$$\begin{aligned} p_{\text{pred}}(y|x_{\text{new}}) &= \int p_w(y|x) p(w|\mathcal{D}) dw \approx \int p_w(y|x) p_{\text{laplace}}(w|\mathcal{D}) dw \\ &= \mathbb{E} \left[\Phi(W^T x_{\text{new}}) \right], \quad W \sim \mathcal{N}(w_{\text{MAP}}, [\nabla^2 E(w_{\text{MAP}})]^{-1}) \\ &= \mathbb{E} [\Phi(G(x_{\text{new}}))] , \quad G(x_{\text{new}}) \sim ?? \text{ 🖊️} . \end{aligned}$$

Model comparison and BIC

- For model comparison, recall that we aim to compute the normalizing constant/marginal likelihood

$$e^{-E(w)} = L(w; \mathcal{D}) p_{\text{prior}}(w), \quad Z(\mathcal{M}) = \int e^{-E(w)} dw.$$

- Here we drop the dependency with respect to \mathcal{M} which specifies p_{prior} or hyperparameters.
- The larger $Z(\mathcal{M})$ is, the better \mathcal{M} is.
- We can apply the same approximation procedure directly to $E(w)$:

$$E(w) \approx E(w_{\text{MAP}}) + \nabla E(w_{\text{MAP}})(w - w_{\text{MAP}}) + (1/2) \nabla^2 E(w_{\text{MAP}})(w - w_{\text{MAP}}).$$

- This gives then that

$$\begin{aligned} \log Z &\approx -E(w_{\text{MAP}}) - (1/2) \log \det(\nabla^2 E(w_{\text{MAP}})) + (d/2) \log(2\pi) \\ &\approx \log L(w_{\text{MAP}}; \mathcal{D}) + \log p_{\text{prior}}(w_{\text{MAP}}) \\ &\quad - (1/2) \log \det(\nabla^2 E(w_{\text{MAP}})) + (d/2) \log(2\pi). \end{aligned}$$

Model comparison and BIC

- For model comparison, recall that we aim to compute the normalizing constant/marginal likelihood


$$e^{-E(w)} = L(w; \mathcal{D}) p_{\text{prior}}(w), \quad Z(\mathcal{M}) = \int e^{-E(w)} dw.$$

- The Laplace approximation is

$$\begin{aligned} \log Z \approx \log L(w_{\text{MAP}}; \mathcal{D}) + \log p_{\text{prior}}(w_{\text{MAP}}) \\ - (1/2) \log \det(\nabla^2 E(w_{\text{MAP}})) + (d/2) \log(2\pi). \end{aligned} \quad (1)$$

- Approximating $\log \det(\nabla^2 E(w_{\text{MAP}})) \approx d \log(N)$ gives rise to the Bayesian Information Criterion (BIC) or the Schwarz criterion:

$$\text{BIC} = \log L(w_{\text{MAP}}; \mathcal{D}) - (d/2) \log(N).$$

-  BIC can also give misleading results.
- In particular, the assumption that the Hessian matrix has full rank is often not valid since many of the parameters are not “well-determined”.
- As possible, use (1)!

- Relation between Fisher linear discriminant and least-square estimation [Bis07, Section 4.1.5];
- 4.3.6 Canonical link functions

- [Bis07] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. 1st ed. Springer, 2007. ISBN: 0387310738.