# MAP 534

# Introduction to machine learning

## Linear models for classification II

Alain Durmus

# Outline

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,

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

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

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

  and aim to quantify if possible the uncertainties of our predictions.
- Here $\mathscr{C}$ is called a classifier.
- Recall that the three paradigms are:
    - deterministic discriminative learning ✓;
    - probabilistic generative learning ✓;
    - probabilistic discriminative learning $\implies$ today;
- We will use this three paradigms to learn particular classifiers.
- A first question we deal with is the existence of an optimal classifier, also called Bayes classifier.

## 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

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

- 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
$$\mathrm{L}(w;\mathcal{D}) = \prod_{i=1}^{N} \{p_w(y_i|x_i) p_X(x_i)\} \,.$$
- 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 $p_{\hat{w}}(y|x_{\mathbf{new}})$ and the prediction:
$$y_{\mathbf{pred}} = \operatorname*{argmax}_y p_{\hat{w}}(y|x_{\mathbf{new}}) \,.$$
- ? 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

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

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

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

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

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

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

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

**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:

$$p_w(1|x_{\mathsf{new}}) = \sigma(a(x_{\mathsf{new}}, w))\,, \quad \sigma(t) = \frac{\mathrm{e}^t}{1+\mathrm{e}^t}$$

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

- The activation $a(x_{\mathsf{new}}, \hat{w})$ = log odds/ log 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 $a(x_{\mathsf{new}}, w) = \phi(x_{\mathsf{new}})^{\mathrm{T}} w$, linear with respect to the parameter.

# Multiclass discrtiminative 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:
$$p_{\boldsymbol{W}}(k|x_{\text{new}}) = \frac{\exp(a(x_{\text{new}}, w_k))}{\sum_{j=1}^{K} \exp(a(x_{\text{new}}, w_j))} ,$$
where $\boldsymbol{W} = \{w_j\}_{j=1}^{K}$ are the parameter to infer.
- The function $\boldsymbol{\sigma} : \mathbb{R}^K \to \mathrm{Simplex}_K = \{(\varpi_1, \ldots, \varpi_K) \in [0,1]^K \ : \ \sum_{k=1}^{K} \varpi_k = 1\}$:
$$\boldsymbol{\sigma}(a_1, \ldots, a_K) = \left( \frac{a_1}{\sum_{j=1}^{K} a_j}, \ldots, \frac{a_K}{\sum_{j=1}^{K} a_j} \right)^{\mathrm{T}}$$
is called the normalized exponential or softmax function.
- With this notation, $p_{\boldsymbol{W}}(k|x_{\text{new}}) = \boldsymbol{\sigma}(a_1, \ldots, a_K)_k$, with $a_k = a(x_{\text{new}}, w_k)$.
- This remark suggests also to directly choose the function $a : \mathrm{X} \times \mathbb{R}^d \to \mathbb{R}$.

# Outline

Probabilistic discriminative models for classification

Logistic regression

Bayesian logistic regression

- We consider here fixed basis functions $\phi = \{\phi_j\}_{j=1}^{d}$, $\phi_j : \mathsf{X} \to \mathbb{R}$.
- 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 $x_i \leftarrow \{\phi_1(x_i), \ldots, \phi_d(x_i)\}$.
- However, we will see in our next lecture that it is much more efficient to consider basis functions which are adaptive and learned from the data.

# Logistic regression: likelihood

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

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

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

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

- The likelihood is then given by

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

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

# Logistic regression: likelihood

- We first consider the $0-1$ prediction task, $\mathrm{Y}=\{0,1\}$.
- We consider here the family of conditional/predictive distributions:

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

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

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

- The likelihood is then given by

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

- Note that the distribution on $X$ is here not relevant.
- Estimator for $w$ by maximum likelihood procedure.

# MLE for the logistic regression

- Maximizing $\mathrm{L}$ is equivalent to minimizing the negative log-likelihood which gives rise to the error function:

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

- $E$ is the called the cross-entropy error function.
- The gradient of $E$ is given by

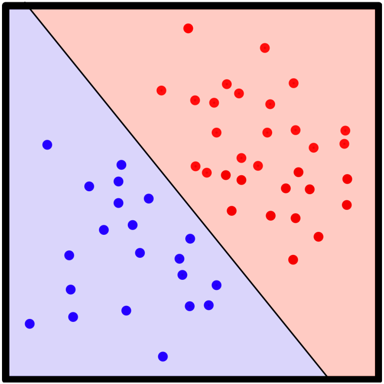
$$\nabla E(w) = \sum_{i=1}^{N} \{y_i - \sigma_i(w)\} x_i . \tag{1}$$

- The proof for (1) uses

$$\frac{\mathrm{d}\sigma}{\mathrm{d}a}(a) = \sigma(a)(1 - \sigma(a)) , \quad \nabla_w \sigma_i(w) = \sigma_i(w)(1 - \sigma_i(w)) x_i .$$

- 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

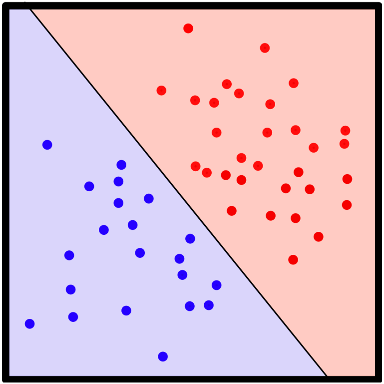


- In the case the data is linearly separable: i.e., there exists $\bar{w} \in \mathbb{R}^d$ such that

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

  then MLE for logistic regression exhibits severe overfitting.
- MLE tends to provide parameters $\hat{w} = \alpha \bar{w}$, $\alpha \in \mathbb{R}$, with $|\alpha|$ very large.
- This results in predictive $p_{\hat{w}}(1|x_{\mathsf{new}})$ which tends to the non-smooth Heaviside function: $\mathbb{1}\left\{\bar{w}^{\mathrm{T}} x < 0\right\}$.
- Problem remains even as $N$ is large.

# Overfitting phenomenon for the logistic regression

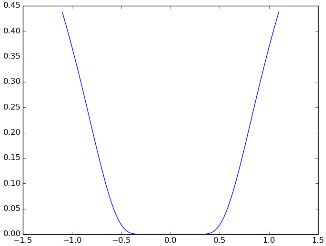


- In the case the data is linearly separable: i.e., there exists $\bar{w} \in \mathbb{R}^d$ such that

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

  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 $E$:

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

## Convexity of the logistic regression and vanishing gradient phenomenon

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

**Proof.**

$\square$

- However, $\min \mathrm{Spec} \nabla^2 E(w) \approx 0$ 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 information.
- Example ✎.

# Iterated reweighted least square/ Newton algorithm

- Newton algorithm aim 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 aim 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 $\mathsf{Y} = \{1, \ldots K\}$, $K > 2$.
- Recall that we start with the statistical model:

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

where $\sigma$ is an activation function.

- This generalizes easily as

$$Y \sim \mathrm{Multi}(\boldsymbol{\sigma}_1(\boldsymbol{W}, X), \ldots, \boldsymbol{\sigma}_K(\boldsymbol{W}, X)) , \quad \boldsymbol{W} = [w_1 \cdots w_K] \in \mathbb{R}^{d \times K} ,$$

where

$$\boldsymbol{\sigma} : \mathbb{R}^{d \times K} \times \mathsf{X} \to \mathrm{Simplex}_K = \{(\varpi_1, \ldots, \varpi_K) \in [0,1]^K \; : \; \sum_{k=1}^{K} \varpi_k = 1\} .$$

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

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

$$Y \sim \mathrm{Multi}(\boldsymbol{\sigma}_1(W, X), \ldots, \boldsymbol{\sigma}_K(W, X)) , \quad W = [w_1 \cdots w_K] \in \mathbb{R}^{d \times K} ,$$

where
$\boldsymbol{\sigma} : \mathbb{R}^{d\times K} \times \mathrm{X} \to \mathrm{Simplex}_K = \{(\varpi_1, \ldots, \varpi_K) \in [0,1]^K \ : \ \sum_{k=1}^K \varpi_k = 1\}$,

$$\boldsymbol{\sigma}_k(W, X) = \mathrm{softmax}(X^{\mathrm{T}} w_1, \ldots, X^{\mathrm{T}} w_K)_k = \frac{\exp(X^{\mathrm{T}} w_k)}{\sum_{j'=1}^K \exp(X^{\mathrm{T}} w_{j'})} .$$

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

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

setting

$$\sigma_{k,i} = \boldsymbol{\sigma}_k(W, x_i) .$$

# Multiclass logistic regression

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

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

  setting

$$\sigma_{k,i} = \boldsymbol{\sigma}_k(\boldsymbol{W}, x_i) .$$

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

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

- ? $\nabla_{\boldsymbol{W}} E$ ✎.
- ? $\nabla^2_{\boldsymbol{W}} E$ ✎.

# Probit regression

- We focus on a $\{0,1\}$ classification task and consider a discreminative probabilistic model:
$$Y \sim \mathrm{Ber}(\sigma(w^{\mathrm{T}} X)) \ ,$$
where $\sigma : \mathbb{R} \to [0,1]$ 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 $\sigma$...
- The probit regression just consists in taking $\sigma = \Phi$ where $\Phi$ is the cumulative distribution function associated with $\mathrm{N}(0,1)$:
$$\Phi(t) = \int_{-\infty}^{t} \mathrm{e}^{-u^2/2} \mathrm{d}u/(2\pi)^{1/2} \ .$$
- 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...

# Outline

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 $w$, $p_{\text{prior}}$, we get using the Bayes formula with the posterior distribution

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

- 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 $\mathbb{R}^d$

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

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

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

- Theoretical justification: Berstein von Mises theorem shows that an appropriate scaled version of the posterior converges to a Gaussian distribution.
- ? which $\mu, \Sigma$?
- We consider for $\mu$ the MAP estimator:

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

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

# Laplace approximation method

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

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

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

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

- We consider for $\mu$ the MAP estimator:

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

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

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

- This suggests to take

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

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

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

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

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

- We consider for $\mu$ the MAP estimator:

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

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

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

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

# Model comparison and BIC

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

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

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

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

- This gives then that

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

# Model comparison and BIC

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

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

- The Laplace approximation is

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

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

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

- $\mathrm{BIC}$ 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 (2)!

- Recall that the posterior has a form

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

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

where

$$\sigma_i = ??$$

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

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

- Finally the Laplace approximation is

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

# Predictive distribution

- The Laplace approximation is

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

- The predictive distribution is itself approximated by

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

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

- Recall that the posterior has a form

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

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

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

- Finally the Laplace approximation is

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

# Predictive distribution

- The Laplace approximation is

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

- The predictive distribution is itself approximated by

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

## Further reading

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

# Bibliography i

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