Bayesian Logistic Regression. Laplace Approximation

Evgeny Burnaev

Skoltech, Moscow, Russia





Burnaev, Bayesian ML Skoltech

Outline

- Bayesian Linear Models for Classification
- 2 Laplace Approximation
- Bayesian Logistic Regression
- 4 Relevance Vector Machine for Classification
- **5** RVM application examples

- Bayesian Linear Models for Classification
- 2 Laplace Approximation
- Bayesian Logistic Regression
- Relevance Vector Machine for Classification
- **6** RVM application examples

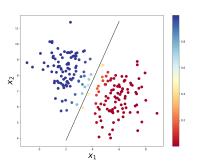
Skoltech
Suilable freizige at Science and Technology 3/41

- We consider a two-class classification problem with classes C_0 and C_1 .
- A logistic model definies i.i.d. probability to obtain a particular class given vector of inputs x:

$$p(C_1|\boldsymbol{\phi}) = t(\boldsymbol{\phi}) = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^\top), \ p(C_0|\boldsymbol{\phi}) = 1 - p(C_1|\boldsymbol{\phi}).$$

Here $\phi = \phi(\mathbf{x})$ is a vector of basis functions

• A data set $\mathcal{D}_m = \{(m{\phi}_i, y_i)\}_{i=1}^m$, where $y_i \in \{0, 1\}$ and $m{\phi}_i = \phi(\mathbf{x}_i)$



(ㅁ▶◀♬▶◀불▶◀불▶ 불 쒸٩♡

Burnaev, Bayesian ML SKOITECN

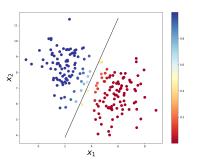
SAME TO THE PROPERTY OF THE PRO

- We consider a two-class classification problem with classes C_0 and C_1 .
- A logistic model definies i.i.d. probability to obtain a particular class given vector of inputs x:

$$p(C_1|\boldsymbol{\phi}) = t(\boldsymbol{\phi}) = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^\top), \ p(C_0|\boldsymbol{\phi}) = 1 - p(C_1|\boldsymbol{\phi}).$$

Here $\phi = \phi(\mathbf{x})$ is a vector of basis functions

• A data set $\mathcal{D}_m = \{(\phi_i, y_i)\}_{i=1}^m$, where $y_i \in \{0, 1\}$ and $\phi_i = \phi(\mathbf{x}_i)$



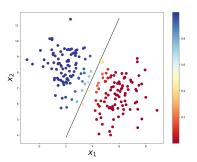
《□》 《□》 《臣》 《臣》 臣 釣@@

- We consider a two-class classification problem with classes C_0 and C_1 .
- A logistic model definies i.i.d. probability to obtain a particular class given vector of inputs x:

$$p(C_1|\boldsymbol{\phi}) = t(\boldsymbol{\phi}) = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^\top), \ p(C_0|\boldsymbol{\phi}) = 1 - p(C_1|\boldsymbol{\phi}).$$

Here $\phi = \phi(\mathbf{x})$ is a vector of basis functions

• A data set $\mathcal{D}_m = \{(m{\phi}_i, y_i)\}_{i=1}^m$, where $y_i \in \{0, 1\}$ and $m{\phi}_i = \phi(\mathbf{x}_i)$



(ロ) (레) (토) (토) 토 9Q@

- ullet We consider a two-class classification problem with classes \mathcal{C}_0 and \mathcal{C}_1 .
- A logistic model definies i.i.d. probability to obtain a particular class given vector of inputs x:

$$p(C_1|\boldsymbol{\phi}) = t(\boldsymbol{\phi}) = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^\top), \ p(C_0|\boldsymbol{\phi}) = 1 - p(C_1|\boldsymbol{\phi}).$$

Here $\phi = \phi(\mathbf{x})$ is a vector of basis functions

- A data set $\mathcal{D}_m = \{(\phi_i, y_i)\}_{i=1}^m$, where $y_i \in \{0, 1\}$ and $\phi_i = \phi(\mathbf{x}_i)$
- The likelihood $p(\mathbf{Y}_m|\mathbf{w})$:

5/41

- ullet We consider a two-class classification problem with classes \mathcal{C}_0 and \mathcal{C}_1 .
- A logistic model definies i.i.d. probability to obtain a particular class given vector of inputs x:

$$p(C_1|\boldsymbol{\phi}) = t(\boldsymbol{\phi}) = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^\top), \ p(C_0|\boldsymbol{\phi}) = 1 - p(C_1|\boldsymbol{\phi}).$$

Here $\phi = \phi(\mathbf{x})$ is a vector of basis functions

- A data set $\mathcal{D}_m=\{(m{\phi}_i,y_i)\}_{i=1}^m$, where $y_i\in\{0,1\}$ and $m{\phi}_i=\phi(\mathbf{x}_i)$
- The likelihood $p(\mathbf{Y}_m|\mathbf{w})$:

$$p(\mathbf{Y}_m|\mathbf{w}) = \prod_{i=1}^m t_i^{y_i} (1-t_i)^{1-y_i},$$

$$\mathbf{Y}_{m} = (y_{1}, \dots, y_{m})^{\top}, \ t_{i} = p(\mathcal{C}_{1} | \phi_{i}).$$

4□ > <</p>
4□ >
4 = >
5
9
0

5/41

$$E(\mathbf{w}) = -\log p(\mathbf{Y}_m | \mathbf{w}) = -\sum_{i=1}^{m} \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\},$$

where $t_i = \sigma(a_i)$ and $a_i = \mathbf{w} \cdot \boldsymbol{\phi}_i^{\top}$.

We easily get that

$$\nabla E(\mathbf{w}) = \sum_{i=1}^{m} (t_i - y_i) \phi_i$$

as

$$\frac{\partial \sigma(a)}{\partial a} = \sigma(a) - \sigma(a)^2$$

(ロ) (部) (注) (注) 注 り(0)

$$E(\mathbf{w}) = -\log p(\mathbf{Y}_m | \mathbf{w}) = -\sum_{i=1}^{m} \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\},\$$

where $t_i = \sigma(a_i)$ and $a_i = \mathbf{w} \cdot \boldsymbol{\phi}_i^{\top}$.

We easily get that

$$\nabla E(\mathbf{w}) = \sum_{i=1}^{m} (t_i - y_i) \phi_i,$$

as

$$\frac{\partial \sigma(a)}{\partial a} = \sigma(a) - \sigma(a)^2.$$

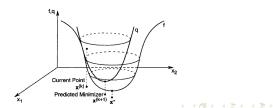
SKOITECH halloon frattate of Science and Technology 6/41

- Newton-Raphson method to minimizes a twice continuously differentiable function $f(\mathbf{x}): \mathbb{R}^d \to \mathbb{R}$

$$f(\mathbf{x}) \approx q(\mathbf{x}) = f(\mathbf{x}^{(k)}) + (\mathbf{x} - \mathbf{x}^{(k)})^{\top} \mathbf{g}^{(k)} + \frac{1}{2} (\mathbf{x} - \mathbf{x}^{(k)})^{\top} F(\mathbf{x}^{(k)}) (\mathbf{x} - \mathbf{x}^{(k)}),$$

$$\mathbf{0} = \nabla q(\mathbf{x}) = \mathbf{g}^{(k)} + F(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)})$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - F(\mathbf{x}^{(k)})^{-1} \mathbf{g}^{(k)}$$



- Newton-Raphson method to minimizes a twice continuously differentiable function $f(\mathbf{x}):\mathbb{R}^d\to\mathbb{R}$
- ullet Obtain a quadratic approximation about the current point ${f x}^{(k)}$

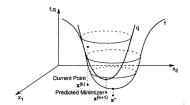
$$f(\mathbf{x}) \approx q(\mathbf{x}) = f(\mathbf{x}^{(k)}) + (\mathbf{x} - \mathbf{x}^{(k)})^{\top} \mathbf{g}^{(k)} + \frac{1}{2} (\mathbf{x} - \mathbf{x}^{(k)})^{\top} F(\mathbf{x}^{(k)}) (\mathbf{x} - \mathbf{x}^{(k)}),$$
where $\mathbf{g}^{(k)} = \nabla f(\mathbf{x}^{(k)}), F(\mathbf{x}^{(k)}) = \nabla \nabla f(\mathbf{x}^{(k)})$

ullet Applying first-order necessary condition to q we get

$$0 = \nabla q(\mathbf{x}) = \mathbf{g}^{(k)} + F(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)})$$

• If $F(\mathbf{x}^{(k)}) > 0$, then $q(\mathbf{x})$ achieves minimum at

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - F(\mathbf{x}^{(k)})^{-1}\mathbf{g}^{(k)}$$



) P P ・ E ト (目) ・ (I) ・ (I)

7/41

- Newton-Raphson method to minimizes a twice continuously differentiable function $f(\mathbf{x}):\mathbb{R}^d\to\mathbb{R}$
- ullet Obtain a quadratic approximation about the current point ${f x}^{(k)}$

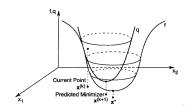
$$f(\mathbf{x}) \approx q(\mathbf{x}) = f(\mathbf{x}^{(k)}) + (\mathbf{x} - \mathbf{x}^{(k)})^{\top} \mathbf{g}^{(k)} + \frac{1}{2} (\mathbf{x} - \mathbf{x}^{(k)})^{\top} F(\mathbf{x}^{(k)}) (\mathbf{x} - \mathbf{x}^{(k)}),$$
 where $\mathbf{g}^{(k)} = \nabla f(\mathbf{x}^{(k)})$, $F(\mathbf{x}^{(k)}) = \nabla \nabla f(\mathbf{x}^{(k)})$

• Applying first-order necessary condition to q we get

$$\mathbf{0} = \nabla q(\mathbf{x}) = \mathbf{g}^{(k)} + F(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)})$$

• If $F(\mathbf{x}^{(k)}) > 0$, then $q(\mathbf{x})$ achieves minimum at

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - F(\mathbf{x}^{(k)})^{-1}\mathbf{g}^{(k)}$$



- Newton-Raphson method to minimizes a twice continuously differentiable function $f(\mathbf{x}):\mathbb{R}^d\to\mathbb{R}$
- ullet Obtain a quadratic approximation about the current point ${f x}^{(k)}$

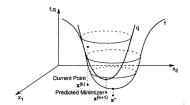
$$f(\mathbf{x}) \approx q(\mathbf{x}) = f(\mathbf{x}^{(k)}) + (\mathbf{x} - \mathbf{x}^{(k)})^{\top} \mathbf{g}^{(k)} + \frac{1}{2} (\mathbf{x} - \mathbf{x}^{(k)})^{\top} F(\mathbf{x}^{(k)}) (\mathbf{x} - \mathbf{x}^{(k)}),$$
 where $\mathbf{g}^{(k)} = \nabla f(\mathbf{x}^{(k)}), F(\mathbf{x}^{(k)}) = \nabla \nabla f(\mathbf{x}^{(k)})$

• Applying first-order necessary condition to q we get

$$\mathbf{0} = \nabla q(\mathbf{x}) = \mathbf{g}^{(k)} + F(\mathbf{x}^{(k)})(\mathbf{x} - \mathbf{x}^{(k)})$$

ullet If $F(\mathbf{x}^{(k)})>0$, then $q(\mathbf{x})$ achieves minimum at

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - F(\mathbf{x}^{(k)})^{-1}\mathbf{g}^{(k)}$$



Ckoltoob

Newton-Raphson method

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - \mathbf{H}^{-1} \nabla E(\mathbf{w})$$

• In case of the linear regression with sum-of-squares error and $m \times M$ design matrix ${\bf \Phi}$

• For a Newtown-Raphson method we get

SKOITECH Salbono Frotzure al Science and Technology 8

Newton-Raphson method

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - \mathbf{H}^{-1} \nabla E(\mathbf{w})$$

• In case of the linear regression with sum-of-squares error and $m \times M$ design matrix ${\bf \Phi}$

For a Newtown-Raphson method we get

◆ロ > ◆ 個 > ◆ 達 > ◆ 達 ・ 釣 へ ②

Newton-Raphson method

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - \mathbf{H}^{-1} \nabla E(\mathbf{w})$$

 \bullet In case of the linear regression with sum-of-squares error and $m\times M$ design matrix ${\bf \Phi}$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} (y_i - \mathbf{w} \cdot \boldsymbol{\phi}_i^{\top})^2,$$

For a Newtown-Raphson method we get

◆ロ > ◆ 個 > ◆ 達 > ◆ 達 ・ 釣 へ ②

8/41

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - \mathbf{H}^{-1} \nabla E(\mathbf{w})$$

• In case of the linear regression with sum-of-squares error and $m \times M$ design matrix ${\bf \Phi}$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} (y_i - \mathbf{w} \cdot \boldsymbol{\phi}_i^{\top})^2,$$

$$\nabla E(\mathbf{w}) = \sum_{i=1}^{m} (\mathbf{w} \cdot \boldsymbol{\phi}_i^{\top} - y_i) \boldsymbol{\phi}_i = \boldsymbol{\Phi}^{\top} \boldsymbol{\Phi} \mathbf{w} - \boldsymbol{\Phi}^{\top} \mathbf{Y}_m,$$

For a Newtown-Raphson method we get

<ロ > ◆ □ > ◆ □ > ◆ □ > ● ● 9 へ ○

8/41

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - \mathbf{H}^{-1} \nabla E(\mathbf{w})$$

 \bullet In case of the linear regression with sum-of-squares error and $m\times M$ design matrix ${\bf \Phi}$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} (y_i - \mathbf{w} \cdot \boldsymbol{\phi}_i^{\top})^2,$$

$$\nabla E(\mathbf{w}) = \sum_{i=1}^{m} (\mathbf{w} \cdot \boldsymbol{\phi}_i^{\top} - y_i) \boldsymbol{\phi}_i = \boldsymbol{\varPhi}^{\top} \boldsymbol{\varPhi} \mathbf{w} - \boldsymbol{\varPhi}^{\top} \mathbf{Y}_m,$$

$$\mathbf{H} = \nabla \nabla E(\mathbf{w}) = \sum_{i=1}^{m} \boldsymbol{\phi}_i^{\top} \cdot \boldsymbol{\phi}_i = \boldsymbol{\varPhi}^{\top} \boldsymbol{\varPhi}.$$

For a Newtown-Raphson method we get

4□ > <</p>
4□ >
4 = >
5
9
0

8/41

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - \mathbf{H}^{-1} \nabla E(\mathbf{w})$$

 \bullet In case of the linear regression with sum-of-squares error and $m\times M$ design matrix ${\bf \Phi}$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} (y_i - \mathbf{w} \cdot \boldsymbol{\phi}_i^{\top})^2,$$

$$\nabla E(\mathbf{w}) = \sum_{i=1}^{m} (\mathbf{w} \cdot \boldsymbol{\phi}_i^{\top} - y_i) \boldsymbol{\phi}_i = \boldsymbol{\varPhi}^{\top} \boldsymbol{\varPhi} \mathbf{w} - \boldsymbol{\varPhi}^{\top} \mathbf{Y}_m,$$

$$\mathbf{H} = \nabla \nabla E(\mathbf{w}) = \sum_{i=1}^{m} \boldsymbol{\phi}_i^{\top} \cdot \boldsymbol{\phi}_i = \boldsymbol{\varPhi}^{\top} \boldsymbol{\varPhi}.$$

• For a Newtown-Raphson method we get

◆ロ > ◆部 > ◆注 > ・注 ・ りへで

8/41

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - \mathbf{H}^{-1} \nabla E(\mathbf{w})$$

• In case of the linear regression with sum-of-squares error and $m \times M$ design matrix ${\bf \Phi}$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} (y_i - \mathbf{w} \cdot \boldsymbol{\phi}_i^{\top})^2,$$

$$\nabla E(\mathbf{w}) = \sum_{i=1}^{m} (\mathbf{w} \cdot \boldsymbol{\phi}_i^{\top} - y_i) \boldsymbol{\phi}_i = \boldsymbol{\varPhi}^{\top} \boldsymbol{\varPhi} \mathbf{w} - \boldsymbol{\varPhi}^{\top} \mathbf{Y}_m,$$

$$\mathbf{H} = \nabla \nabla E(\mathbf{w}) = \sum_{i=1}^{m} \boldsymbol{\phi}_i^{\top} \cdot \boldsymbol{\phi}_i = \boldsymbol{\varPhi}^{\top} \boldsymbol{\varPhi}.$$

• For a Newtown-Raphson method we get

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - (\boldsymbol{\varPhi}^{\top}\boldsymbol{\varPhi})^{-1} \left(\boldsymbol{\varPhi}^{\top}\boldsymbol{\varPhi}\mathbf{w}^{(old)} - \boldsymbol{\varPhi}\mathbf{Y}_{m}\right)$$

8/41

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - \mathbf{H}^{-1} \nabla E(\mathbf{w})$$

• In case of the linear regression with sum-of-squares error and $m \times M$ design matrix ${\bf \Phi}$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} (y_i - \mathbf{w} \cdot \boldsymbol{\phi}_i^{\top})^2,$$

$$\nabla E(\mathbf{w}) = \sum_{i=1}^{m} (\mathbf{w} \cdot \boldsymbol{\phi}_i^{\top} - y_i) \boldsymbol{\phi}_i = \boldsymbol{\varPhi}^{\top} \boldsymbol{\varPhi} \mathbf{w} - \boldsymbol{\varPhi}^{\top} \mathbf{Y}_m,$$

$$\mathbf{H} = \nabla \nabla E(\mathbf{w}) = \sum_{i=1}^{m} \boldsymbol{\phi}_i^{\top} \cdot \boldsymbol{\phi}_i = \boldsymbol{\varPhi}^{\top} \boldsymbol{\varPhi}.$$

• For a Newtown-Raphson method we get

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - (\boldsymbol{\Phi}^{\top} \boldsymbol{\Phi})^{-1} \left(\boldsymbol{\Phi}^{\top} \boldsymbol{\Phi} \mathbf{w}^{(old)} - \boldsymbol{\Phi} \mathbf{Y}_{m} \right)$$
$$= (\boldsymbol{\Phi}^{\top} \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^{\top} \mathbf{Y}_{m}.$$

《□》《문》《불》 불 / 잇()

$$E(\mathbf{w}) = -\log p(\mathbf{Y}_m | \mathbf{w}) = -\sum_{i=1}^{m} \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\},\$$

where $t_i = \sigma(a_i)$ and $a_i = \mathbf{w} \cdot \boldsymbol{\phi}_i^{\top}$.

Newton-Raphson method for logistic regression

Since $0 < t_i < 1$, then $\mathbf{z}^{\top} \mathbf{H} \mathbf{z} > 0$, i.e. \mathbf{H} is positive definite. Thus the error function is a concave function.

◆□ → ◆部 → ◆差 → を りゅ

$$E(\mathbf{w}) = -\log p(\mathbf{Y}_m | \mathbf{w}) = -\sum_{i=1}^{m} \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\},\$$

where $t_i = \sigma(a_i)$ and $a_i = \mathbf{w} \cdot \boldsymbol{\phi}_i^{\top}$.

Newton-Raphson method for logistic regression

Since $0 < t_i < 1$, then $\mathbf{z}^{\top} \mathbf{H} \mathbf{z} > 0$, i.e. \mathbf{H} is positive definite. Thus the error function is a concave function.

9/41

$$E(\mathbf{w}) = -\log p(\mathbf{Y}_m | \mathbf{w}) = -\sum_{i=1}^{m} \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\},\$$

where $t_i = \sigma(a_i)$ and $a_i = \mathbf{w} \cdot \boldsymbol{\phi}_i^{\top}$.

Newton-Raphson method for logistic regression

$$\nabla E(\mathbf{w}) = \sum_{i=1}^{m} (t_i - y_i) \phi_i = \mathbf{\Phi}^{\top} (\mathbf{t} - \mathbf{Y}_m),$$

Since $0 < t_i < 1$, then $\mathbf{z}^{\top} \mathbf{H} \mathbf{z} > 0$, i.e. \mathbf{H} is positive definite. Thus the error function is a concave function.

9/41

$$E(\mathbf{w}) = -\log p(\mathbf{Y}_m | \mathbf{w}) = -\sum_{i=1}^{m} \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\},\$$

where $t_i = \sigma(a_i)$ and $a_i = \mathbf{w} \cdot \boldsymbol{\phi}_i^{\top}$.

Newton-Raphson method for logistic regression

$$\nabla E(\mathbf{w}) = \sum_{i=1}^{m} (t_i - y_i) \boldsymbol{\phi}_i = \boldsymbol{\Phi}^{\top} (\mathbf{t} - \mathbf{Y}_m),$$
$$\mathbf{H} = \nabla \nabla E(\mathbf{w}) = \sum_{i=1}^{m} t_i (1 - t_i) \boldsymbol{\phi}_n^{\top} \boldsymbol{\phi}_n = \boldsymbol{\Phi}^{\top} \mathbf{R} \boldsymbol{\Phi}.$$

• Since $0 < t_i < 1$, then $\mathbf{z}^{\top} \mathbf{H} \mathbf{z} > 0$, i.e. \mathbf{H} is positive definite. Thus the error function is a concave function.

$$E(\mathbf{w}) = -\log p(\mathbf{Y}_m | \mathbf{w}) = -\sum_{i=1}^{m} \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\},\$$

where $t_i = \sigma(a_i)$ and $a_i = \mathbf{w} \cdot \boldsymbol{\phi}_i^{\top}$.

Newton-Raphson method for logistic regression

$$\nabla E(\mathbf{w}) = \sum_{i=1}^{m} (t_i - y_i) \boldsymbol{\phi}_i = \boldsymbol{\Phi}^{\top} (\mathbf{t} - \mathbf{Y}_m),$$
$$\mathbf{H} = \nabla \nabla E(\mathbf{w}) = \sum_{i=1}^{m} t_i (1 - t_i) \boldsymbol{\phi}_n^{\top} \boldsymbol{\phi}_n = \boldsymbol{\Phi}^{\top} \mathbf{R} \boldsymbol{\Phi}.$$

Here \mathbf{R} is a diagonal matrix $m \times m$ with elements $R_{ii} = t_i(1 - t_i)$

• Since $0 < t_i < 1$, then $\mathbf{z}^{\top} \mathbf{H} \mathbf{z} > 0$, i.e. \mathbf{H} is positive definite. Thus the error function is a concave function.

9/41

$$E(\mathbf{w}) = -\log p(\mathbf{Y}_m | \mathbf{w}) = -\sum_{i=1}^{m} \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\},\$$

where $t_i = \sigma(a_i)$ and $a_i = \mathbf{w} \cdot \boldsymbol{\phi}_i^{\top}$.

Newton-Raphson method for logistic regression

$$\nabla E(\mathbf{w}) = \sum_{i=1}^{m} (t_i - y_i) \boldsymbol{\phi}_i = \boldsymbol{\Phi}^{\top} (\mathbf{t} - \mathbf{Y}_m),$$
$$\mathbf{H} = \nabla \nabla E(\mathbf{w}) = \sum_{i=1}^{m} t_i (1 - t_i) \boldsymbol{\phi}_n^{\top} \boldsymbol{\phi}_n = \boldsymbol{\Phi}^{\top} \mathbf{R} \boldsymbol{\Phi}.$$

Here ${\bf R}$ is a diagonal matrix $m \times m$ with elements $R_{ii} = t_i(1-t_i)$

• Since $0 < t_i < 1$, then $\mathbf{z}^{\top} \mathbf{H} \mathbf{z} > 0$, i.e. \mathbf{H} is positive definite. Thus the error function is a concave function.

4 D > 4 D > 4 E > 4 E > E 9 Q C

Burnaev, Bayesian ML

9/41

• R can be interpreted as a covariance matrix, since

$$\mathbb{E}[y_i] = \sigma(a_i) = t_i,$$

$$\text{var}[y_i] = \mathbb{E}[y_i^2] - (\mathbb{E}[y_i])^2 = \sigma(a_i) - \sigma^2(a_i) = t_i(1 - t_i) = [\mathbf{R}]_{ii}.$$

Skoltech

10/41

$$\mathbf{w}^{(new)} = \mathbf{w}^{(old)} - (\mathbf{\varPhi}^{\top}\mathbf{R}\mathbf{\varPhi})^{-1}\mathbf{\varPhi}^{\top}(\mathbf{t} - \mathbf{Y}_m)$$

• R can be interpreted as a covariance matrix, since

$$\mathbb{E}[y_i] = \sigma(a_i) = t_i,$$

$$\text{var}[y_i] = \mathbb{E}[y_i^2] - (\mathbb{E}[y_i])^2 = \sigma(a_i) - \sigma^2(a_i) = t_i(1 - t_i) = [\mathbf{R}]_{ii}.$$

<ロ > ◆ □ > ◆ □ > ◆ □ > ● ● 9 へ ○

Skoltech
Suitous instan et Science and Technology 10/41

$$\begin{aligned} \mathbf{w}^{(new)} &= \mathbf{w}^{(old)} - (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \boldsymbol{\varPhi}^{\top} (\mathbf{t} - \mathbf{Y}_m) \\ &= (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \left\{ \boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi} \mathbf{w}^{(old)} - \boldsymbol{\varPhi}^{\top} (\mathbf{t} - \mathbf{Y}_m) \right\} \end{aligned}$$

• R can be interpreted as a covariance matrix, since

$$\mathbb{E}[y_i] = \sigma(a_i) = t_i,$$

$$\text{var}[y_i] = \mathbb{E}[y_i^2] - (\mathbb{E}[y_i])^2 = \sigma(a_i) - \sigma^2(a_i) = t_i(1 - t_i) = [\mathbf{R}]_{ii}.$$

<ロ > ◆ □ > ◆ □ > ◆ □ > ● ● 9 へ ○

Skoltech

10/41

$$\begin{split} \mathbf{w}^{(new)} &= \mathbf{w}^{(old)} - (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \boldsymbol{\varPhi}^{\top} (\mathbf{t} - \mathbf{Y}_m) \\ &= (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \left\{ \boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi} \mathbf{w}^{(old)} - \boldsymbol{\varPhi}^{\top} (\mathbf{t} - \mathbf{Y}_m) \right\} \\ &= (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \boldsymbol{\varPhi}^{\top} \mathbf{R} \left\{ \boldsymbol{\varPhi} \mathbf{w}^{(old)} - \mathbf{R}^{-1} (\mathbf{t} - \mathbf{Y}_m) \right\} \end{split}$$

• R can be interpreted as a covariance matrix, since

$$\mathbb{E}[y_i] = \sigma(a_i) = t_i,$$

$$var[y_i] = \mathbb{E}[y_i^2] - (\mathbb{E}[y_i])^2 = \sigma(a_i) - \sigma^2(a_i) = t_i(1 - t_i) = [\mathbf{R}]_{ii}.$$

Skoltech

$$\begin{split} \mathbf{w}^{(new)} &= \mathbf{w}^{(old)} - (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \boldsymbol{\varPhi}^{\top} (\mathbf{t} - \mathbf{Y}_m) \\ &= (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \left\{ \boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi} \mathbf{w}^{(old)} - \boldsymbol{\varPhi}^{\top} (\mathbf{t} - \mathbf{Y}_m) \right\} \\ &= (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \boldsymbol{\varPhi}^{\top} \mathbf{R} \left\{ \boldsymbol{\varPhi} \mathbf{w}^{(old)} - \mathbf{R}^{-1} (\mathbf{t} - \mathbf{Y}_m) \right\} \\ &= (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \boldsymbol{\varPhi}^{\top} \mathbf{R} \mathbf{z}. \end{split}$$

R can be interpreted as a covariance matrix, since

$$\mathbb{E}[y_i] = \sigma(a_i) = t_i,$$

$$\text{var}[y_i] = \mathbb{E}[y_i^2] - (\mathbb{E}[y_i])^2 = \sigma(a_i) - \sigma^2(a_i) = t_i(1 - t_i) = [\mathbf{R}]_{ii}.$$

4 D > 4 B > 4 B > B 9 9 9

Skoltech
Stations Insitate of Science and Technology 10/41

$$\begin{split} \mathbf{w}^{(new)} &= \mathbf{w}^{(old)} - (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \boldsymbol{\varPhi}^{\top} (\mathbf{t} - \mathbf{Y}_m) \\ &= (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \left\{ \boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi} \mathbf{w}^{(old)} - \boldsymbol{\varPhi}^{\top} (\mathbf{t} - \mathbf{Y}_m) \right\} \\ &= (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \boldsymbol{\varPhi}^{\top} \mathbf{R} \left\{ \boldsymbol{\varPhi} \mathbf{w}^{(old)} - \mathbf{R}^{-1} (\mathbf{t} - \mathbf{Y}_m) \right\} \\ &= (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \boldsymbol{\varPhi}^{\top} \mathbf{R} \mathbf{z}. \end{split}$$

Here

$$\mathbf{z} = \boldsymbol{\Phi} \mathbf{w}^{(old)} - \mathbf{R}^{-1} (\mathbf{t} - \mathbf{Y}_m).$$

• R can be interpreted as a covariance matrix, since

$$\mathbb{E}[y_i] = \sigma(a_i) = t_i,$$

$$\text{var}[y_i] = \mathbb{E}[y_i^2] - (\mathbb{E}[y_i])^2 = \sigma(a_i) - \sigma^2(a_i) = t_i(1 - t_i) = [\mathbf{R}]_{ii}.$$

Skoltech

$$\begin{split} \mathbf{w}^{(new)} &= \mathbf{w}^{(old)} - (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \boldsymbol{\varPhi}^{\top} (\mathbf{t} - \mathbf{Y}_m) \\ &= (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \left\{ \boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi} \mathbf{w}^{(old)} - \boldsymbol{\varPhi}^{\top} (\mathbf{t} - \mathbf{Y}_m) \right\} \\ &= (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \boldsymbol{\varPhi}^{\top} \mathbf{R} \left\{ \boldsymbol{\varPhi} \mathbf{w}^{(old)} - \mathbf{R}^{-1} (\mathbf{t} - \mathbf{Y}_m) \right\} \\ &= (\boldsymbol{\varPhi}^{\top} \mathbf{R} \boldsymbol{\varPhi})^{-1} \boldsymbol{\varPhi}^{\top} \mathbf{R} \mathbf{z}. \end{split}$$

Here

$$\mathbf{z} = \mathbf{\Phi} \mathbf{w}^{(old)} - \mathbf{R}^{-1} (\mathbf{t} - \mathbf{Y}_m).$$

• R can be interpreted as a covariance matrix, since

$$\mathbb{E}[y_i] = \sigma(a_i) = t_i, \text{var}[y_i] = \mathbb{E}[y_i^2] - (\mathbb{E}[y_i])^2 = \sigma(a_i) - \sigma^2(a_i) = t_i(1 - t_i) = [\mathbf{R}]_{ii}.$$

4 D F 4 A F F 4 B F

Probit regression

• The general model

$$p(t=1|a) = f(a), \ a = \mathbf{w} \cdot \boldsymbol{\phi}^{\top},$$

- $f(\cdot)$ is an activation function:
 - $-t_i=1$, if $a_i\geq \theta$
 - $t_i = 0$, otherwise
- Usually we consider the noisy threshold model $\theta \sim p(\theta)$, $f(a) = \int_{-\infty}^{a} p(\theta) d\theta$.
- For the logit (logistic) regression $f(a) = \sigma(a)$ and a logistic distribution as $p(\theta)$.
- For the probit regression

Skoltech

11/41

The general model

$$p(t=1|a) = f(a), \ a = \mathbf{w} \cdot \boldsymbol{\phi}^{\top},$$

- $f(\cdot)$ is an activation function:
 - $-t_i=1$, if $a_i\geq \theta$
- $-t_i=0$, otherwise
- Usually we consider the noisy threshold model $\theta \sim p(\theta)$, $f(a) = \int_{-\infty}^{a} p(\theta) d\theta$.
- For the logit (logistic) regression $f(a) = \sigma(a)$ and a logistic distribution as $p(\theta)$.
- For the probit regression

The general model

$$p(t=1|a) = f(a), \ a = \mathbf{w} \cdot \boldsymbol{\phi}^{\top},$$

- $f(\cdot)$ is an activation function:
 - $-t_i=1$, if $a_i\geq \theta$
- $-t_i=0$, otherwise
- Usually we consider the noisy threshold model $\theta \sim p(\theta)$, $f(a) = \int_{-\infty}^{a} p(\theta) d\theta$.
- For the logit (logistic) regression $f(a) = \sigma(a)$ and a logistic distribution as $p(\theta)$.
- For the probit regression

Skoltech

11/41

The general model

$$p(t=1|a) = f(a), \ a = \mathbf{w} \cdot \boldsymbol{\phi}^{\top},$$

- $f(\cdot)$ is an activation function:
 - $-t_i=1$, if $a_i \geq \theta$
- $-t_i=0$, otherwise
- Usually we consider the noisy threshold model $\theta \sim p(\theta)$, $f(a) = \int_{-\infty}^{a} p(\theta) d\theta$.
- For the logit (logistic) regression $f(a) = \sigma(a)$ and a logistic distribution as $p(\theta)$.
- For the probit regression

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0, 1) d\theta$$

The general model

$$p(t=1|a) = f(a), \ a = \mathbf{w} \cdot \boldsymbol{\phi}^{\top},$$

 $f(\cdot)$ is an activation function:

- $-t_i=1$, if $a_i \geq \theta$
- $-t_i=0$, otherwise
- Usually we consider the noisy threshold model $\theta \sim p(\theta)$, $f(a) = \int_{-\infty}^a p(\theta) d\theta$.
- For the logit (logistic) regression $f(a) = \sigma(a)$ and a logistic distribution as $p(\theta)$.
- For the probit regression

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0, 1)d\theta$$

$$\operatorname{erf}(a) = \frac{2}{\sqrt{\pi}} \int_{0}^{a} \exp(-\theta^{2}/2)d\theta$$

Burnaev, Bayesian ML Subservative Concess through 11/41

The general model

$$p(t=1|a) = f(a), \ a = \mathbf{w} \cdot \boldsymbol{\phi}^{\top},$$

 $f(\cdot)$ is an activation function:

- $-t_i=1$, if $a_i\geq \theta$
- $-t_i=0$, otherwise
- Usually we consider the noisy threshold model $\theta \sim p(\theta)$, $f(a) = \int_{-\infty}^a p(\theta) d\theta$.
- For the logit (logistic) regression $f(a) = \sigma(a)$ and a logistic distribution as $p(\theta)$.
- For the probit regression

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0, 1)d\theta$$

$$\operatorname{erf}(a) = \frac{2}{\sqrt{\pi}} \int_{0}^{a} \exp(-\theta^{2}/2)d\theta$$

$$f(a) = \Phi(a) = \frac{1}{2} \left\{ 1 + \frac{1}{\sqrt{2}} \operatorname{erf}(a) \right\}$$

4 D > 4 B > 4 B > B 9 9 9

Southone Protitage of Science and Technology 111/41

The general model

$$p(t=1|a) = f(a), \ a = \mathbf{w} \cdot \boldsymbol{\phi}^{\top},$$

- $f(\cdot)$ is an activation function:
 - $-t_i=1$, if $a_i \geq \theta$
- $-t_i=0$, otherwise
- Usually we consider the noisy threshold model $\theta \sim p(\theta)$, $f(a) = \int_{-\infty}^a p(\theta) d\theta$.
- For the logit (logistic) regression $f(a) = \sigma(a)$ and a logistic distribution as $p(\theta)$.
- For the probit regression

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0, 1)d\theta$$

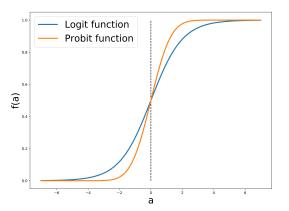
$$\operatorname{erf}(a) = \frac{2}{\sqrt{\pi}} \int_{0}^{a} \exp(-\theta^{2}/2)d\theta$$

$$f(a) = \Phi(a) = \frac{1}{2} \left\{ 1 + \frac{1}{\sqrt{2}} \operatorname{erf}(a) \right\}$$

$$p(\mathcal{C}_{1}|\phi) = t(\phi) = \Phi(\mathbf{w} \cdot \phi^{\top})$$

Difference between the logit and the probit regression

- They are not that much different.
- But the probit is a more Bayes-friendly option (more on this later!).



- Bayesian Linear Models for Classification
- 2 Laplace Approximation
- 3 Bayesian Logistic Regression
- 4 Relevance Vector Machine for Classification
- 5 RVM application examples

Approximate inference

- We have a distribution p(z).
- This distribution p(z)
 - is too complex
 - is known only up to a normalization constant
 - doesn't integrate analytically (almost every Bayesian inference problem).
- We want to find q(z) that is in some sense close to p(z), but is better from our point of view (e.g. the integral is tractable):

$$q(z) \approx p(z)$$

Skoltech
Suzkone tracker at Science and Technology 14/41

Approximate inference

- We have a distribution p(z).
- This distribution p(z)
 - is too complex
 - is known only up to a normalization constant
 - doesn't integrate analytically (almost every Bayesian inference problem).
- We want to find q(z) that is in some sense close to p(z), but is better from our point of view (e.g. the integral is tractable):

$$q(z) \approx p(z)$$

Approximate inference

- We have a distribution p(z).
- This distribution p(z)
 - is too complex
 - is known only up to a normalization constant
 - doesn't integrate analytically (almost every Bayesian inference problem).
- We want to find q(z) that is in some sense close to p(z), but is better from our point of view (e.g. the integral is tractable):

$$q(z) \approx p(z)$$
.

Skoltech
Skuldooo Institute of Science and Technology 14/41

Density with an unknown normalization constant

$$p(z) = \frac{1}{Z}f(z), \quad Z = \int f(z)dz.$$

We calculate a mode of the distribution

$$p'(z_0) = 0 \Leftrightarrow \frac{df(z)}{dz}\big|_{z=z_0} = 0$$

Using the Taylor approximation we get that

$$\log f(z) \approx \log f(z_0) - \frac{1}{2}A(z - z_0)^2, \ A = -\frac{d^2}{dz^2}\log f(z)\Big|_{z=z_0}$$

Thus we get that

$$f(z) \approx f(z_0) \exp \left\{ -\frac{A}{2} (z - z_0)^2 \right\}$$

◆□ → ◆□ → ◆■ → ◆■ → ◆○○

Skoltech
Stutions Institute of Science and Technology 15/41

Density with an unknown normalization constant

$$p(z) = \frac{1}{Z}f(z), \quad Z = \int f(z)dz.$$

We calculate a mode of the distribution

$$p'(z_0) = 0 \Leftrightarrow \frac{df(z)}{dz}\big|_{z=z_0} = 0$$

Using the Taylor approximation we get that

$$\log f(z) \approx \log f(z_0) - \frac{1}{2}A(z - z_0)^2, \ A = -\frac{d^2}{dz^2}\log f(z)\Big|_{z=z_0}$$

Thus we get that

$$f(z) \approx f(z_0) \exp \left\{ -\frac{A}{2} (z - z_0)^2 \right\}$$

(ㅁ▶◀畵▶◀불▶◀불▶ 불 쒸٩♡

Burnaev, Bayesian ML Subscript State of the state of the

Density with an unknown normalization constant

$$p(z) = \frac{1}{Z}f(z), \quad Z = \int f(z)dz.$$

We calculate a mode of the distribution

$$p'(z_0) = 0 \Leftrightarrow \frac{df(z)}{dz}\big|_{z=z_0} = 0$$

Using the Taylor approximation we get that

$$\log f(z) \approx \log f(z_0) - \frac{1}{2}A(z - z_0)^2, \ A = -\frac{d^2}{dz^2}\log f(z)\Big|_{z=z_0}$$

Thus we get that

$$f(z) \approx f(z_0) \exp\left\{-\frac{A}{2}(z - z_0)^2\right\}$$

4 D > 4 B > 4 B > B 9 Q G

Density with an unknown normalization constant

$$p(z) = \frac{1}{Z}f(z), \quad Z = \int f(z)dz.$$

We calculate a mode of the distribution

$$p'(z_0) = 0 \Leftrightarrow \frac{df(z)}{dz}\big|_{z=z_0} = 0$$

Using the Taylor approximation we get that

$$\log f(z) \approx \log f(z_0) - \frac{1}{2}A(z - z_0)^2, \ A = -\frac{d^2}{dz^2}\log f(z)\Big|_{z=z_0}$$

Thus we get that

$$f(z) \approx f(z_0) \exp\left\{-\frac{A}{2}(z-z_0)^2\right\}$$

15/41

• The normalization constant

$$Z = \int f(z)dz \approx f(z_0) \int \exp\left\{-\frac{A}{2}(z - z_0)^2\right\} dz = f(z_0) \left(\frac{2\pi}{A}\right)^{1/2}$$

Thus we get that

Institute of Science and Technology 16/41

• The normalization constant

$$Z = \int f(z)dz \approx f(z_0) \int \exp\left\{-\frac{A}{2}(z - z_0)^2\right\} dz = f(z_0) \left(\frac{2\pi}{A}\right)^{1/2}$$

Thus we get that

Stations institute at Science and Technology 16/41

• The normalization constant

$$Z = \int f(z)dz \approx f(z_0) \int \exp\left\{-\frac{A}{2}(z - z_0)^2\right\} dz = f(z_0) \left(\frac{2\pi}{A}\right)^{1/2}$$

Thus we get that

$$p(z) = \frac{f(z)}{Z} \approx \frac{f(z_0) \exp\left\{-\frac{A}{2}(z - z_0)^2\right\}}{f(z_0) \left(\frac{2\pi}{A}\right)^{1/2}}$$

Skoltech 16/

The normalization constant

$$Z = \int f(z)dz \approx f(z_0) \int \exp\left\{-\frac{A}{2}(z - z_0)^2\right\} dz = f(z_0) \left(\frac{2\pi}{A}\right)^{1/2}$$

Thus we get that

$$p(z) = \frac{f(z)}{Z} \approx \frac{f(z_0) \exp\left\{-\frac{A}{2}(z - z_0)^2\right\}}{f(z_0) \left(\frac{2\pi}{A}\right)^{1/2}}$$
$$= \left(\frac{A}{2\pi}\right)^{1/2} \exp\left\{-\frac{A}{2}(z - z_0)^2\right\}$$

Burnaev, Bayesian ML Superior of Circles as I forward 16/41

Laplace Approximation is an approximation by Gaussian distribution

We got the density

$$p(z) \approx q(z) = \left(\frac{A}{2\pi}\right)^{1/2} \exp\left\{-\frac{A}{2}(z-z_0)^2\right\}.$$

• It is the density of Gaussian distribution

$$q(z) = \mathcal{N}(z|z_0, A^{-1}).$$



Burnaev, Bayesian ML 17/41

Laplace Approximation is an approximation by Gaussian distribution

We got the density

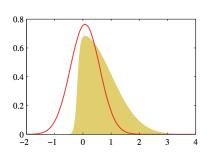
$$p(z) \approx q(z) = \left(\frac{A}{2\pi}\right)^{1/2} \exp\left\{-\frac{A}{2}(z-z_0)^2\right\}.$$

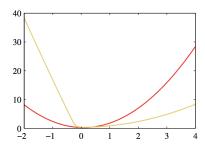
• It is the density of Gaussian distribution:

$$q(z) = \mathcal{N}(z|z_0, A^{-1}).$$



Burnaev, Bayesian ML 2004 Construction of the Construction of the



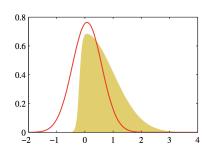


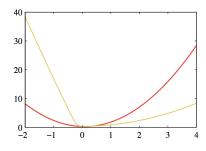
- Laplace approximation for $p(z) \sim \exp(-z^2/2)\sigma(20z+4)$
- The left plot: the normalized distribution p(z) in yellow, the Laplace approximation centred on the mode z_0 of p(z) in rec
- The right plot: the negative logarithms of the corresponding curves

| ㅁㅏㅓ@ㅏㅓㅌㅏㅓㅌㅏ | ㅌ | 쒸٩@

Skoltech

18/41

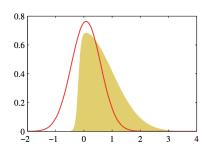


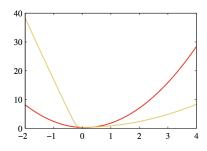


- Laplace approximation for $p(z) \sim \exp(-z^2/2)\sigma(20z+4)$
- \bullet The left plot: the normalized distribution p(z) in yellow, the Laplace approximation centred on the mode z_0 of p(z) in red
- The right plot: the negative logarithms of the corresponding curves

(ㅁㅏㅓ@ㅏㅓㅌㅏㅓㅌㅏ = = ~)٩()

Burnaev, Bayesian ML Skoltech





- Laplace approximation for $p(z) \sim \exp(-z^2/2)\sigma(20z+4)$
- \bullet The left plot: the normalized distribution p(z) in yellow, the Laplace approximation centred on the mode z_0 of p(z) in red
- \bullet The right plot: the negative logarithms of the corresponding curves

18/41

Burnaev, Bayesian ML Subset Indiana of Science and Science

Density with unknown normalization constant

$$p(\mathbf{z}) = f(\mathbf{z})/Z, \ Z = \int f(\mathbf{z})d\mathbf{z}$$

Taylor expansion

$$\log f(\mathbf{z}) \approx \log f(\mathbf{z}_0) - (\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) / 2$$
$$\mathbf{A} = -\nabla \nabla \log f(\mathbf{z}) \Big|_{\mathbf{z} = \mathbf{z}_0}$$

We get

$$f(\mathbf{z}) \approx f(\mathbf{z}_0) \exp\left\{-(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0)/2\right\}$$

Normalization constant

Density with unknown normalization constant

$$p(\mathbf{z}) = f(\mathbf{z})/Z, \ Z = \int f(\mathbf{z})d\mathbf{z}$$

Taylor expansion

$$\log f(\mathbf{z}) \approx \log f(\mathbf{z}_0) - (\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) / 2,$$
$$\mathbf{A} = -\nabla \nabla \log f(\mathbf{z}) \Big|_{\mathbf{z} = \mathbf{z}_0}$$

We get

$$f(\mathbf{z}) \approx f(\mathbf{z}_0) \exp\left\{-(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0)/2\right\}$$

Normalization constant

noticate at Science and Technology 19/41

Density with unknown normalization constant

$$p(\mathbf{z}) = f(\mathbf{z})/Z, \ Z = \int f(\mathbf{z})d\mathbf{z}$$

Taylor expansion

$$\log f(\mathbf{z}) \approx \log f(\mathbf{z}_0) - (\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) / 2,$$
$$\mathbf{A} = -\nabla \nabla \log f(\mathbf{z}) \Big|_{\mathbf{z} = \mathbf{z}_0}$$

We get

$$f(\mathbf{z}) \approx f(\mathbf{z}_0) \exp\left\{-(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0)/2\right\}$$

Normalization constant

Density with unknown normalization constant

$$p(\mathbf{z}) = f(\mathbf{z})/Z, \ Z = \int f(\mathbf{z})d\mathbf{z}$$

Taylor expansion

$$\log f(\mathbf{z}) \approx \log f(\mathbf{z}_0) - (\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) / 2,$$
$$\mathbf{A} = -\nabla \nabla \log f(\mathbf{z}) \Big|_{\mathbf{z} = \mathbf{z}_0}$$

We get

$$f(\mathbf{z}) \approx f(\mathbf{z}_0) \exp\left\{-(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0)/2\right\}$$

Normalization constant

19/41

Density with unknown normalization constant

$$p(\mathbf{z}) = f(\mathbf{z})/Z, \ Z = \int f(\mathbf{z})d\mathbf{z}$$

Taylor expansion

$$\log f(\mathbf{z}) \approx \log f(\mathbf{z}_0) - (\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) / 2,$$
$$\mathbf{A} = -\nabla \nabla \log f(\mathbf{z}) \Big|_{\mathbf{z} = \mathbf{z}_0}$$

We get

$$f(\mathbf{z}) \approx f(\mathbf{z}_0) \exp\left\{-(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0)/2\right\}$$

Normalization constant

$$Z = \int f(\mathbf{z}) d\mathbf{z}$$

skorteen 19/41

Density with unknown normalization constant

$$p(\mathbf{z}) = f(\mathbf{z})/Z, \ Z = \int f(\mathbf{z})d\mathbf{z}$$

Taylor expansion

$$\log f(\mathbf{z}) \approx \log f(\mathbf{z}_0) - (\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) / 2,$$
$$\mathbf{A} = -\nabla \nabla \log f(\mathbf{z}) \Big|_{\mathbf{z} = \mathbf{z}_0}$$

We get

$$f(\mathbf{z}) \approx f(\mathbf{z}_0) \exp\left\{-(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0)/2\right\}$$

Normalization constant

$$Z = \int f(\mathbf{z}) d\mathbf{z}$$

$$\approx f(\mathbf{z}_0) \int \exp\{-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A}(\mathbf{z} - \mathbf{z}_0)\} d\mathbf{z}$$

(ロ) (B) (토) (토) (토) (Q)

Burnaev, Bayesian ML Substitute of Service and Inchession 19/41

Density with unknown normalization constant

$$p(\mathbf{z}) = f(\mathbf{z})/Z, \ Z = \int f(\mathbf{z})d\mathbf{z}$$

Taylor expansion

$$\log f(\mathbf{z}) \approx \log f(\mathbf{z}_0) - (\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) / 2,$$
$$\mathbf{A} = -\nabla \nabla \log f(\mathbf{z}) \Big|_{\mathbf{z} = \mathbf{z}_0}$$

We get

$$f(\mathbf{z}) \approx f(\mathbf{z}_0) \exp\left\{-(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0)/2\right\}$$

Normalization constant

$$Z = \int f(\mathbf{z}) d\mathbf{z}$$

$$\approx f(\mathbf{z}_0) \int \exp\{-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A}(\mathbf{z} - \mathbf{z}_0)\} d\mathbf{z}$$

$$= f(\mathbf{z}_0) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}$$

Station instant of Science and Technology

Multidimensional Laplace Approximation

Laplace approximation has the form

$$p(\mathbf{z}) = \frac{f(\mathbf{z})}{Z}$$

$$\mathbf{z}_0 = \arg\max_{\mathbf{z}} f(\mathbf{z}),$$

$$\mathbf{z}_0 = \arg \max_{\mathbf{z}} f(\mathbf{z}),$$

$$\mathbf{A} = -\nabla \nabla \log f(\mathbf{z}) \Big|_{\mathbf{z} = \mathbf{z}_0}$$

<ロティ部ティミティミテー語

Burnaev, Bayesian ML

20/41

Laplace approximation has the form

$$p(\mathbf{z}) = \frac{f(\mathbf{z})}{Z}$$

$$\approx \frac{f(\mathbf{z}_0) \exp\left\{-(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0)/2\right\}}{f(\mathbf{z}_0) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}}$$

Here

$$\mathbf{z}_0 = \arg \max_{\mathbf{z}} f(\mathbf{z}),$$

$$\mathbf{A} = -\nabla \nabla \log f(\mathbf{z}) \Big|_{\mathbf{z} = \mathbf{z}_0}$$

It is again a Gaussian distribution

(ロ) (部) (注) (注) 注 り(0)

Laplace approximation has the form

$$p(\mathbf{z}) = \frac{f(\mathbf{z})}{Z}$$

$$\approx \frac{f(\mathbf{z}_0) \exp\left\{-(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) / 2\right\}}{f(\mathbf{z}_0) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}}$$

$$= \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\frac{1}{2} (\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0)\right\} = \mathcal{N}(\mathbf{z}|\mathbf{z}_0, \mathbf{A}^{-1})$$

Here

$$\mathbf{z}_0 = \arg \max_{\mathbf{z}} f(\mathbf{z}),$$

$$\mathbf{A} = -\nabla \nabla \log f(\mathbf{z}) \Big|_{\mathbf{z} = \mathbf{z}_0}$$

It is again a Gaussian distribution

(□ → ∢御 → ∢産 → ∢産 → 、差 …のQ@

Skoltech Skilon broken et Science and Technology Laplace approximation has the form

$$p(\mathbf{z}) = \frac{f(\mathbf{z})}{Z}$$

$$\approx \frac{f(\mathbf{z}_0) \exp\left\{-(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0) / 2\right\}}{f(\mathbf{z}_0) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}}$$

$$= \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\frac{1}{2} (\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0)\right\} = \mathcal{N}(\mathbf{z}|\mathbf{z}_0, \mathbf{A}^{-1})$$

Here

$$\mathbf{z}_0 = \arg\max_{\mathbf{z}} f(\mathbf{z}),$$

$$\mathbf{A} = -\nabla\nabla \log f(\mathbf{z})\Big|_{\mathbf{z} = \mathbf{z}_0}$$

• It is again a Gaussian distribution

4 ロ ト 4 回 ト 4 重 ト 4 重 ト 9 9 9 9

Burnaev, Bayesian ML Substitute of Substitut

- Bayesian Linear Models for Classification
- 2 Laplace Approximation
- 3 Bayesian Logistic Regression
- 4 Relevance Vector Machine for Classification
- **(5)** RVM application examples

Prior over parameters

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_0, \mathbf{S}_0)$$

Probability is

$$t_i = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}_i^\top)$$

Bayes formula

$$\mathbf{Y}_m = (y_1, \dots, y_m)^{\top}, \ p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$

Log-posterior

• Maximizing $\log p(\mathbf{w}|\mathbf{Y}_m)$ we estimate \mathbf{w}_{MAP}

Prior over parameters

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_0, \mathbf{S}_0)$$

Probability is

$$t_i = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}_i^\top)$$

Bayes formula

$$\mathbf{Y}_m = (y_1, \dots, y_m)^{\top}, \ p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$

Log-posterior

• Maximizing $\log p(\mathbf{w}|\mathbf{Y}_m)$ we estimate \mathbf{w}_{MAP}

Bayesian Logistic Regression

Prior over parameters

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_0, \mathbf{S}_0)$$

Probability is

$$t_i = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}_i^\top)$$

Bayes formula

$$\mathbf{Y}_m = (y_1, \dots, y_m)^{\top}, \ p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$

Log-posterior

• Maximizing $\log p(\mathbf{w}|\mathbf{Y}_m)$ we estimate \mathbf{w}_{MAP}

Bayesian Logistic Regression

Prior over parameters

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_0, \mathbf{S}_0)$$

Probability is

$$t_i = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}_i^{\top})$$

Bayes formula

$$\mathbf{Y}_m = (y_1, \dots, y_m)^{\top}, \ p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$

Log-posterior

ullet Maximizing $\log p(\mathbf{w}|\mathbf{Y}_m)$ we estimate \mathbf{w}_{MAP}

Bayesian Logistic Regression

Prior over parameters

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_0, \mathbf{S}_0)$$

Probability is

$$t_i = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}_i^\top)$$

Bayes formula

$$\mathbf{Y}_m = (y_1, \dots, y_m)^{\top}, \ p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$

Log-posterior

$$\log p(\mathbf{w}|\mathbf{Y}_m) = -\frac{1}{2}(\mathbf{w} - \boldsymbol{\omega}_0)^{\top} \mathbf{S}_0^{-1}(\mathbf{w} - \boldsymbol{\omega}_0) +$$

ullet Maximizing $\log p(\mathbf{w}|\mathbf{Y}_m)$ we estimate \mathbf{w}_{MAP}

◆ロ > ← 個 > ← 速 > ← 速 > ・ 速 ・ り へ で

Skoltech

22/41

Prior over parameters

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_0, \mathbf{S}_0)$$

Probability is

$$t_i = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}_i^\top)$$

Bayes formula

$$\mathbf{Y}_m = (y_1, \dots, y_m)^{\mathsf{T}}, \ p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$

Log-posterior

$$\log p(\mathbf{w}|\mathbf{Y}_m) = -\frac{1}{2}(\mathbf{w} - \boldsymbol{\omega}_0)^{\top} \mathbf{S}_0^{-1}(\mathbf{w} - \boldsymbol{\omega}_0) + \sum_{i=1}^{m} \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\} + \text{const} \to \max_{\mathbf{w}}$$

• Maximizing $\log p(\mathbf{w}|\mathbf{Y}_m)$ we estimate \mathbf{w}_{MAP}

4 D > 4 B > 4 B > B 9 Q Q

Prior over parameters

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_0, \mathbf{S}_0)$$

Probability is

$$t_i = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}_i^\top)$$

Bayes formula

$$\mathbf{Y}_m = (y_1, \dots, y_m)^{\top}, \ p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$

Log-posterior

$$\log p(\mathbf{w}|\mathbf{Y}_m) = -\frac{1}{2}(\mathbf{w} - \boldsymbol{\omega}_0)^{\top} \mathbf{S}_0^{-1}(\mathbf{w} - \boldsymbol{\omega}_0) + \sum_{i=1}^{m} \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\} + \text{const} \to \max_{\mathbf{w}}$$

• Maximizing $\log p(\mathbf{w}|\mathbf{Y}_m)$ we estimate \mathbf{w}_{MAP}

22/41

Burnaev, Bayesian ML

$$\mathbf{Y}_m = (y_1, \dots, y_m)^{\top}, \ p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$

- Gaussian approximation to the posterior
 - first we maximize the posterior to get the MAP
 - second we estimate the Hessian to get covariance matrix

• Laplace (Gaussian) approximation to the posterior $p(\mathbf{w}|\mathbf{Y}_m)$ has the form

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{MAP}, \mathbf{S}_m)$$

(□) (□) (□) (□) (□)

$$\mathbf{Y}_m = (y_1, \dots, y_m)^{\top}, \ p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$

$$\log p(\mathbf{w}|\mathbf{Y}_m) = -\frac{1}{2}(\mathbf{w} - \boldsymbol{\omega}_0)^{\top} \mathbf{S}_0^{-1}(\mathbf{w} - \boldsymbol{\omega}_0) +$$

- Gaussian approximation to the posterior.
 - first we maximize the posterior to get the MAP
 - second we estimate the Hessian to get covariance matrix

• Laplace (Gaussian) approximation to the posterior $p(\mathbf{w}|\mathbf{Y}_m)$ has the form

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{MAP}, \mathbf{S}_m)$$

- 4 ロ ト 4 御 ト 4 恵 ト 4 恵 ト - 恵 - 90

$$\mathbf{Y}_m = (y_1, \dots, y_m)^\top, \quad p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$
$$\log p(\mathbf{w}|\mathbf{Y}_m) = -\frac{1}{2}(\mathbf{w} - \boldsymbol{\omega}_0)^\top \mathbf{S}_0^{-1}(\mathbf{w} - \boldsymbol{\omega}_0) + \sum_{i=1}^m \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\} + \text{const}$$

Gaussian approximation to the posterior:

first we maximize the posterior to get the MAP
second we estimate the Hessian to get covariance n

• Laplace (Gaussian) approximation to the posterior $p(\mathbf{w}|\mathbf{Y}_m)$ has the form

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{MAP}, \mathbf{S}_m)$$

→□▶→□▶→□▶→□ 例

$$\mathbf{Y}_m = (y_1, \dots, y_m)^\top, \ p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$
$$\log p(\mathbf{w}|\mathbf{Y}_m) = -\frac{1}{2}(\mathbf{w} - \boldsymbol{\omega}_0)^\top \mathbf{S}_0^{-1}(\mathbf{w} - \boldsymbol{\omega}_0) + \sum_{i=1}^m \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\} + \text{const}$$

- Gaussian approximation to the posterior:
 - first we maximize the posterior to get the MAP
 - second we estimate the Hessian to get covariance matrix

• Laplace (Gaussian) approximation to the posterior $p(\mathbf{w}|\mathbf{Y}_m)$ has the form

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{MAP}, \mathbf{S}_m)$$

citate al Science and Technology 23/41

$$\mathbf{Y}_m = (y_1, \dots, y_m)^\top, \ p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$
$$\log p(\mathbf{w}|\mathbf{Y}_m) = -\frac{1}{2}(\mathbf{w} - \boldsymbol{\omega}_0)^\top \mathbf{S}_0^{-1}(\mathbf{w} - \boldsymbol{\omega}_0) + \sum_{i=1}^m \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\} + \text{const}$$

- Gaussian approximation to the posterior:
 - first we maximize the posterior to get the MAP
 - second we estimate the Hessian to get covariance matrix

• Laplace (Gaussian) approximation to the posterior $p(\mathbf{w}|\mathbf{Y}_m)$ has the form

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{MAP}, \mathbf{S}_m)$$

Stations institute of Science and Technology 23/41

$$\mathbf{Y}_m = (y_1, \dots, y_m)^{\top}, \ p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$
$$\log p(\mathbf{w}|\mathbf{Y}_m) = -\frac{1}{2}(\mathbf{w} - \boldsymbol{\omega}_0)^{\top} \mathbf{S}_0^{-1}(\mathbf{w} - \boldsymbol{\omega}_0) + \sum_{i=1}^m \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\} + \text{const}$$

- Gaussian approximation to the posterior:
 - first we maximize the posterior to get the MAP
 - second we estimate the Hessian to get covariance matrix

• Laplace (Gaussian) approximation to the posterior $p(\mathbf{w}|\mathbf{Y}_m)$ has the form

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{MAP}, \mathbf{S}_m)$$

$$\mathbf{Y}_m = (y_1, \dots, y_m)^\top, \ p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$
$$\log p(\mathbf{w}|\mathbf{Y}_m) = -\frac{1}{2}(\mathbf{w} - \boldsymbol{\omega}_0)^\top \mathbf{S}_0^{-1}(\mathbf{w} - \boldsymbol{\omega}_0) + \sum_{i=1}^m \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\} + \text{const}$$

- Gaussian approximation to the posterior:
 - first we maximize the posterior to get the MAP
 - second we estimate the Hessian to get covariance matrix

$$\mathbf{S}_{m}^{-1} = -\nabla\nabla\log p(\mathbf{w}|\mathbf{Y}_{m}) = \mathbf{S}_{0}^{-1} + \sum_{i=1}^{m} t_{i}(1 - t_{i})\boldsymbol{\phi}_{i} \cdot \boldsymbol{\phi}_{i}^{\top}\Big|_{\mathbf{w} = \mathbf{w}_{MAP}}$$

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{MAP}, \mathbf{S}_m)$$

$$\mathbf{Y}_m = (y_1, \dots, y_m)^{\top}, \ p(\mathbf{w}|\mathbf{Y}_m) \sim p(\mathbf{w})p(\mathbf{Y}_m|\mathbf{w})$$
$$\log p(\mathbf{w}|\mathbf{Y}_m) = -\frac{1}{2}(\mathbf{w} - \boldsymbol{\omega}_0)^{\top} \mathbf{S}_0^{-1}(\mathbf{w} - \boldsymbol{\omega}_0) + \sum_{i=1}^m \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\} + \text{const}$$

- Gaussian approximation to the posterior:
 - first we maximize the posterior to get the MAP
 - second we estimate the Hessian to get covariance matrix

$$\mathbf{S}_{m}^{-1} = -\nabla\nabla\log p(\mathbf{w}|\mathbf{Y}_{m}) = \mathbf{S}_{0}^{-1} + \sum_{i=1}^{m} t_{i}(1 - t_{i})\boldsymbol{\phi}_{i} \cdot \boldsymbol{\phi}_{i}^{\top}\Big|_{\mathbf{w} = \mathbf{w}_{MAP}}$$

• Laplace (Gaussian) approximation to the posterior $p(\mathbf{w}|\mathbf{Y}_m)$ has the form

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{MAP}, \mathbf{S}_m)$$

23/41

- $p(\mathbf{w}|\mathbf{Y}_m)$ is approximated by a Gaussian $q(\mathbf{w})$
- The class probability

$$p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{Y}_m) = \int p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{w}) p(\mathbf{w}|\mathbf{Y}_m) d\mathbf{w} \approx \int \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^\top) q(\mathbf{w}) d\mathbf{w}$$

ullet Let us denote $a = \mathbf{w} \cdot oldsymbol{\phi}^ op$. We represent

$$\sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) \sigma(a) da$$

Here $\delta(\cdot)$ is a delta-function

We get

- ullet $p(\mathbf{w}|\mathbf{Y}_m)$ is approximated by a Gaussian $q(\mathbf{w})$
- The class probability

$$p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{Y}_m) = \int p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{w}) p(\mathbf{w}|\mathbf{Y}_m) d\mathbf{w} \approx \int \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^\top) q(\mathbf{w}) d\mathbf{w}$$

• Let us denote $a = \mathbf{w} \cdot \boldsymbol{\phi}^{\top}$. We represent

$$\sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) \sigma(a) da$$

Here $\delta(\cdot)$ is a delta-function

We get

Skoltech

24/41

- $p(\mathbf{w}|\mathbf{Y}_m)$ is approximated by a Gaussian $q(\mathbf{w})$
- The class probability

$$p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{Y}_m) = \int p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{w}) p(\mathbf{w}|\mathbf{Y}_m) d\mathbf{w} \approx \int \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^\top) q(\mathbf{w}) d\mathbf{w}$$

• Let us denote $a = \mathbf{w} \cdot \boldsymbol{\phi}^{\top}$. We represent

$$\sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) \sigma(a) da$$

Here $\delta(\cdot)$ is a delta-function

We get

- $p(\mathbf{w}|\mathbf{Y}_m)$ is approximated by a Gaussian $q(\mathbf{w})$
- The class probability

$$p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{Y}_m) = \int p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{w}) p(\mathbf{w}|\mathbf{Y}_m) d\mathbf{w} \approx \int \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^\top) q(\mathbf{w}) d\mathbf{w}$$

• Let us denote $a = \mathbf{w} \cdot \boldsymbol{\phi}^{\top}$. We represent

$$\sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) \sigma(a) da$$

Here $\delta(\cdot)$ is a delta-function

We get

Skoltech

24/41

- ullet $p(\mathbf{w}|\mathbf{Y}_m)$ is approximated by a Gaussian $q(\mathbf{w})$
- The class probability

$$p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{Y}_m) = \int p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{w}) p(\mathbf{w}|\mathbf{Y}_m) d\mathbf{w} \approx \int \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^\top) q(\mathbf{w}) d\mathbf{w}$$

• Let us denote $a = \mathbf{w} \cdot \boldsymbol{\phi}^{\top}$. We represent

$$\sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) \sigma(a) da$$

Here $\delta(\cdot)$ is a delta-function

We get

$$\int \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w} =$$

- $p(\mathbf{w}|\mathbf{Y}_m)$ is approximated by a Gaussian $q(\mathbf{w})$
- The class probability

$$p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{Y}_m) = \int p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{w}) p(\mathbf{w}|\mathbf{Y}_m) d\mathbf{w} \approx \int \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^\top) q(\mathbf{w}) d\mathbf{w}$$

• Let us denote $a = \mathbf{w} \cdot \boldsymbol{\phi}^{\top}$. We represent

$$\sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) \sigma(a) da$$

Here $\delta(\cdot)$ is a delta-function

We get

$$\int \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w} = \int \left[\int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w} \right] \sigma(a) da$$

OKOITECH ONE Profitate of Science and Technology 24/41

- ullet $p(\mathbf{w}|\mathbf{Y}_m)$ is approximated by a Gaussian $q(\mathbf{w})$
- The class probability

$$p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{Y}_m) = \int p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{w}) p(\mathbf{w}|\mathbf{Y}_m) d\mathbf{w} \approx \int \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^\top) q(\mathbf{w}) d\mathbf{w}$$

• Let us denote $a = \mathbf{w} \cdot \boldsymbol{\phi}^{\top}$. We represent

$$\sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) \sigma(a) da$$

Here $\delta(\cdot)$ is a delta-function

We get

$$\int \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w} = \int \left[\int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w} \right] \sigma(a) da$$
$$= \int \sigma(a) p(a) da$$

Koltech
on rozzas at Science and Technology 24/41

- $p(\mathbf{w}|\mathbf{Y}_m)$ is approximated by a Gaussian $q(\mathbf{w})$
- The class probability

$$p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{Y}_m) = \int p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{w}) p(\mathbf{w}|\mathbf{Y}_m) d\mathbf{w} \approx \int \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^\top) q(\mathbf{w}) d\mathbf{w}$$

• Let us denote $a = \mathbf{w} \cdot \boldsymbol{\phi}^{\top}$. We represent

$$\sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) \sigma(a) da$$

Here $\delta(\cdot)$ is a delta-function

We get

$$\int \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w} = \int \left[\int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w} \right] \sigma(a) da$$
$$= \int \sigma(a) p(a) da$$
with
$$p(a) = \int \delta(a - \mathbf{w}^{\top} \boldsymbol{\phi}) q(\mathbf{w}) d\mathbf{w}$$

◆ロト ◆部 ト ◆ 恵 ト ◆ 恵 ・ 夕 ♀

tate of Science and Technology 24/41

$$p(a) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w}$$

- \bullet Delta function imposes a linear constraint on w and so forms a marginal distribution by integrating out all directions orthogonal to ϕ
- ullet So p(a) is 1d Gaussian distribution $\mathcal{N}(\mu_a,\sigma_a^2)$ with

$$p(a) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w}$$

- ullet Delta function imposes a linear constraint on ullet and so forms a marginal distribution by integrating out all directions orthogonal to ϕ
- ullet So p(a) is 1d Gaussian distribution $\mathcal{N}(\mu_a,\sigma_a^2)$ with

$$p(a) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w}$$

- \bullet Delta function imposes a linear constraint on w and so forms a marginal distribution by integrating out all directions orthogonal to ϕ
- ullet So p(a) is 1d Gaussian distribution $\mathcal{N}(\mu_a,\sigma_a^2)$ with

$$p(a) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w}$$

- \bullet Delta function imposes a linear constraint on w and so forms a marginal distribution by integrating out all directions orthogonal to ϕ
- \bullet So p(a) is 1d Gaussian distribution $\mathcal{N}(\mu_a,\sigma_a^2)$ with

$$\mu_a = \mathbb{E}[a] = \int p(a)ada = \int \left[\int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^\top)ada\right] q(\mathbf{w})d\mathbf{w}$$

Skoltech

25/41

$$p(a) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w}$$

- \bullet Delta function imposes a linear constraint on w and so forms a marginal distribution by integrating out all directions orthogonal to ϕ
- \bullet So p(a) is 1d Gaussian distribution $\mathcal{N}(\mu_a,\sigma_a^2)$ with

$$\mu_a = \mathbb{E}[a] = \int p(a)ada = \int \left[\int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^\top) ada \right] q(\mathbf{w}) d\mathbf{w}$$
$$= \int \mathbf{w} \cdot \boldsymbol{\phi}^\top q(\mathbf{w}) d\mathbf{w} = \left[\int \mathbf{w} q(\mathbf{w}) d\mathbf{w} \right] \cdot \boldsymbol{\phi}^\top = \mathbf{w}_{MAP} \cdot \boldsymbol{\phi}^\top$$

◆ロ > ◆ 個 > ◆ 達 > ◆ 達 ・ 釣 へ ②

$$p(a) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w}$$

• So p(a) is 1d Gaussian distribution $\mathcal{N}(\mu_a,\sigma_a^2)$ with

Recall

$$p(a) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w}$$

 \bullet So p(a) is 1d Gaussian distribution $\mathcal{N}(\mu_a,\sigma_a^2)$ with

$$p(a) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w}$$

ullet So p(a) is 1d Gaussian distribution $\mathcal{N}(\mu_a,\sigma_a^2)$ with

$$\sigma_a^2 = \operatorname{var}[a] = \int p(a)(a^2 - (\mathbb{E}[a])^2)da$$

$$p(a) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w}$$

ullet So p(a) is 1d Gaussian distribution $\mathcal{N}(\mu_a,\sigma_a^2)$ with

$$\sigma_a^2 = \text{var}[a] = \int p(a)(a^2 - (\mathbb{E}[a])^2)da$$
$$= \int \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^\top) q(\mathbf{w})(a^2 - (\mathbb{E}[a])^2)dad\mathbf{w}$$

$$p(a) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w}$$

• So p(a) is 1d Gaussian distribution $\mathcal{N}(\mu_a, \sigma_a^2)$ with

$$\sigma_a^2 = \text{var}[a] = \int p(a)(a^2 - (\mathbb{E}[a])^2)da$$

$$= \int \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^\top) q(\mathbf{w})(a^2 - (\mathbb{E}[a])^2) da d\mathbf{w}$$

$$= \int \left[\int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^\top) (a^2 - (\mathbb{E}[a])^2) da \right] q(\mathbf{w}) d\mathbf{w}$$

$$p(a) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w}$$

• So p(a) is 1d Gaussian distribution $\mathcal{N}(\mu_a, \sigma_a^2)$ with

$$\begin{split} \sigma_a^2 &= \mathrm{var}[a] = \int p(a)(a^2 - (\mathbb{E}[a])^2) da \\ &= \int \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^\top) q(\mathbf{w}) (a^2 - (\mathbb{E}[a])^2) da d\mathbf{w} \\ &= \int \left[\int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^\top) (a^2 - (\mathbb{E}[a])^2) da \right] q(\mathbf{w}) d\mathbf{w} \\ &= \int \{ (\mathbf{w} \cdot \boldsymbol{\phi}^\top)^2 - (\mathbf{w}_{MAP} \cdot \boldsymbol{\phi}^\top)^2 \} q(\mathbf{w}) d\mathbf{w} \end{split}$$

26/41

$$p(a) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w}$$

 \bullet So p(a) is 1d Gaussian distribution $\mathcal{N}(\mu_a,\sigma_a^2)$ with

$$\begin{split} \sigma_a^2 &= \mathrm{var}[a] = \int p(a)(a^2 - (\mathbb{E}[a])^2) da \\ &= \int \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^\top) q(\mathbf{w}) (a^2 - (\mathbb{E}[a])^2) da d\mathbf{w} \\ &= \int \left[\int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^\top) (a^2 - (\mathbb{E}[a])^2) da \right] q(\mathbf{w}) d\mathbf{w} \\ &= \int \{ (\mathbf{w} \cdot \boldsymbol{\phi}^\top)^2 - (\mathbf{w}_{MAP} \cdot \boldsymbol{\phi}^\top)^2 \} q(\mathbf{w}) d\mathbf{w} \\ &= \int \boldsymbol{\phi}^\top (\mathbf{w} - \mathbf{w}_{MAP})^\top (\mathbf{w} - \mathbf{w}_{MAP}) \boldsymbol{\phi} q(\mathbf{w}) d\mathbf{w} \end{split}$$

visitable of Science and Technology 26/41

$$p(a) = \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^{\top}) q(\mathbf{w}) d\mathbf{w}$$

 \bullet So p(a) is 1d Gaussian distribution $\mathcal{N}(\mu_a,\sigma_a^2)$ with

$$\sigma_a^2 = \operatorname{var}[a] = \int p(a)(a^2 - (\mathbb{E}[a])^2) da$$

$$= \int \int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^\top) q(\mathbf{w}) (a^2 - (\mathbb{E}[a])^2) da d\mathbf{w}$$

$$= \int \left[\int \delta(a - \mathbf{w} \cdot \boldsymbol{\phi}^\top) (a^2 - (\mathbb{E}[a])^2) da \right] q(\mathbf{w}) d\mathbf{w}$$

$$= \int \{ (\mathbf{w} \cdot \boldsymbol{\phi}^\top)^2 - (\mathbf{w}_{MAP} \cdot \boldsymbol{\phi}^\top)^2 \} q(\mathbf{w}) d\mathbf{w}$$

$$= \int \boldsymbol{\phi}^\top (\mathbf{w} - \mathbf{w}_{MAP})^\top (\mathbf{w} - \mathbf{w}_{MAP}) \boldsymbol{\phi} q(\mathbf{w}) d\mathbf{w}$$

$$= \boldsymbol{\phi}^\top \left[\int (\mathbf{w} - \mathbf{w}_{MAP})^\top (\mathbf{w} - \mathbf{w}_{MAP}) q(\mathbf{w}) d\mathbf{w} \right] \boldsymbol{\phi} = \boldsymbol{\phi}^\top \mathbf{S}_m \boldsymbol{\phi}$$

one frontiere of Science and Technology 26/41

$$p(C_1|\mathbf{Y}_m) = \int \sigma(a)p(a)da = \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2)da$$

• We approximate $\sigma(a)$ by $\Phi(\lambda a)$. Here $\lambda^2=\pi/8$ (same slope at the origin)

$$\sigma(a) \approx \Phi(\lambda a)$$

We get that

$$\int \Phi(\lambda a) \mathcal{N}(a|\mu, \sigma^2) da = \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right)$$

 \bullet Applying the approximation $\sigma(a)\approx \varPhi(\lambda a)$ to both sides we get that

$$\int \sigma(a) \mathcal{N}(a|\mu, \sigma^2) da \approx \sigma(\varkappa(\sigma^2)\mu), \ \varkappa(\sigma^2) = (1 + \pi \sigma^2/8)^{-1/2}$$

Finally

$$p(C_1|\boldsymbol{\phi}, \mathbf{Y}_m) = \sigma(\boldsymbol{\varkappa}(\sigma_a^2)\mu_a)$$

27/41

Burnaev, Bayesian ML Skoltech

$$p(C_1|\mathbf{Y}_m) = \int \sigma(a)p(a)da = \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2)da$$

• We approximate $\sigma(a)$ by $\Phi(\lambda a)$. Here $\lambda^2=\pi/8$ (same slope at the origin)

$$\sigma(a) \approx \Phi(\lambda a)$$

We get that

$$\int \Phi(\lambda a) \mathcal{N}(a|\mu, \sigma^2) da = \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right)$$

 \bullet Applying the approximation $\sigma(a)\approx \varPhi(\lambda a)$ to both sides we get that

$$\int \sigma(a) \mathcal{N}(a|\mu, \sigma^2) da \approx \sigma(\varkappa(\sigma^2)\mu), \ \varkappa(\sigma^2) = (1 + \pi \sigma^2/8)^{-1/2}$$

Finally

$$p(\mathcal{C}_1|\boldsymbol{\phi}, \mathbf{Y}_m) = \sigma(\varkappa(\sigma_a^2)\mu_a)$$

◆□▶ ◆圖▶ ◆基▶ ◆基▶ ■ 釣♀

27/41

Burnaev, Bayesian ML Skoltech

$$p(\mathcal{C}_1|\mathbf{Y}_m) = \int \sigma(a)p(a)da = \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2)da$$

• We approximate $\sigma(a)$ by $\Phi(\lambda a)$. Here $\lambda^2=\pi/8$ (same slope at the origin)

$$\sigma(a) \approx \Phi(\lambda a)$$

We get that

$$\int \Phi(\lambda a) \mathcal{N}(a|\mu, \sigma^2) da = \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right)$$

ullet Applying the approximation $\sigma(a)pprox arPhi(\lambda a)$ to both sides we get that

$$\int \sigma(a) \mathcal{N}(a|\mu, \sigma^2) da \approx \sigma(\varkappa(\sigma^2)\mu), \ \varkappa(\sigma^2) = (1 + \pi \sigma^2/8)^{-1/2}$$

Finally

$$p(C_1|\boldsymbol{\phi}, \mathbf{Y}_m) = \sigma(\varkappa(\sigma_a^2)\mu_a)$$

$$p(C_1|\mathbf{Y}_m) = \int \sigma(a)p(a)da = \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2)da$$

• We approximate $\sigma(a)$ by $\Phi(\lambda a)$. Here $\lambda^2=\pi/8$ (same slope at the origin)

$$\sigma(a) \approx \Phi(\lambda a)$$

We get that

$$\int \Phi(\lambda a) \mathcal{N}(a|\mu, \sigma^2) da = \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right)$$

 \bullet Applying the approximation $\sigma(a) \approx \varPhi(\lambda a)$ to both sides we get that

$$\int \sigma(a) \mathcal{N}(a|\mu, \sigma^2) da \approx \sigma(\varkappa(\sigma^2)\mu), \ \varkappa(\sigma^2) = (1 + \pi \sigma^2/8)^{-1/2}$$

Finally

$$p(C_1|\boldsymbol{\phi}, \mathbf{Y}_m) = \sigma(\varkappa(\sigma_a^2)\mu_a)$$

4□ > 4♂ > 4 ≥ > 4 ≥ > ≥ 9 Q (

sone institute of Science and Technology 27/41

$$p(C_1|\mathbf{Y}_m) = \int \sigma(a)p(a)da = \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2)da$$

• We approximate $\sigma(a)$ by $\Phi(\lambda a)$. Here $\lambda^2=\pi/8$ (same slope at the origin)

$$\sigma(a) \approx \Phi(\lambda a)$$

We get that

$$\int \Phi(\lambda a) \mathcal{N}(a|\mu, \sigma^2) da = \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right)$$

 \bullet Applying the approximation $\sigma(a) \approx \varPhi(\lambda a)$ to both sides we get that

$$\int \sigma(a) \mathcal{N}(a|\mu, \sigma^2) da \approx \sigma(\varkappa(\sigma^2)\mu), \ \varkappa(\sigma^2) = (1 + \pi \sigma^2/8)^{-1/2}$$

Finally

$$p(C_1|\boldsymbol{\phi}, \mathbf{Y}_m) = \sigma(\varkappa(\sigma_a^2)\mu_a)$$

◆□ ト ◆□ ト ◆ 恵 ト ◆ 恵 ・ 夕 Q で

- Bayesian Linear Models for Classification
- 2 Laplace Approximation
- 3 Bayesian Logistic Regression
- 4 Relevance Vector Machine for Classification
- 5 RVM application examples

Class probability

$$t(\boldsymbol{\phi}) = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}), \boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{x})$$

Log-posterior for the prior

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \boldsymbol{\alpha}^{-1}), \ \boldsymbol{\alpha} = \text{diag}\{\alpha_1, \dots, \alpha_M\}$$

has the form

Skoltech

29/41

RVM for Classification

Class probability

$$t(\boldsymbol{\phi}) = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}), \boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{x})$$

Log-posterior for the prior

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \boldsymbol{\alpha}^{-1}), \ \boldsymbol{\alpha} = \operatorname{diag}\{\alpha_1, \dots, \alpha_M\}$$

has the form

Skoltech

Class probability

$$t(\boldsymbol{\phi}) = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}), \boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{x})$$

Log-posterior for the prior

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \boldsymbol{\alpha}^{-1}), \ \boldsymbol{\alpha} = \text{diag}\{\alpha_1, \dots, \alpha_M\}$$

has the form

$$\log p(\mathbf{w}|\mathbf{Y}_m, \boldsymbol{\alpha}) = \log\{p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})\} - \log p(\mathbf{Y}_m|\boldsymbol{\alpha})$$

Section return of Service and Technology

29/41

Class probability

$$t(\boldsymbol{\phi}) = \sigma(\mathbf{w} \cdot \boldsymbol{\phi}^{\top}), \boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{x})$$

Log-posterior for the prior

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \boldsymbol{\alpha}^{-1}), \ \boldsymbol{\alpha} = \text{diag}\{\alpha_1, \dots, \alpha_M\}$$

has the form

$$\log p(\mathbf{w}|\mathbf{Y}_m, \boldsymbol{\alpha}) = \log\{p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})\} - \log p(\mathbf{Y}_m|\boldsymbol{\alpha})$$
$$= \sum_{i=1}^{m} \{y_i \log t_i + (1 - y_i) \log(1 - t_i)\} - \mathbf{w}^{\top} \boldsymbol{\alpha} \mathbf{w}/2 + \text{const}$$

Southout Institute of Science and Technology 2

Due to the Laplace approximation

$$\int_{\mathbf{r}} f(\mathbf{z}) d\mathbf{z} \approx f(\mathbf{z}_0) \int_{\mathbf{r}} \exp\left\{-\frac{1}{2} (\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A} (\mathbf{z} - \mathbf{z}_0)\right\} d\mathbf{z} = f(\mathbf{z}_0) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}},$$

where

$$\mathbf{z}_0 = \arg \max_{\mathbf{z}} f(\mathbf{z}), \ \mathbf{A} = -\nabla \nabla \log f(\mathbf{z}) \Big|_{\mathbf{z} = \mathbf{z}_0}$$

We get that

$$\nabla \log p(\mathbf{w}|\mathbf{Y}_m, \boldsymbol{\alpha}) = \boldsymbol{\Phi}^{\top}(\mathbf{Y}_m - \mathbf{t}) - \boldsymbol{\alpha}\mathbf{w},$$
$$\nabla \nabla \log p(\mathbf{w}|\mathbf{Y}_m, \boldsymbol{\alpha}) = -(\boldsymbol{\Phi}^{\top}\mathbf{R}\boldsymbol{\Phi} + \boldsymbol{\alpha}).$$

where ${f R}$ is an m imes m diagonal matrix with elements $R_{ii}=t_i(1-t_i)$

We get that MAP estimate fulfils equality

$$\mathbf{\Phi}^{\top}(\mathbf{Y}_m - \mathbf{t}) - \mathbf{A}\mathbf{w} = 0 \Rightarrow$$

We can use an approximation to get an estimate w*

$$\mathbf{w}^* = \mathbf{A}^{-1} \boldsymbol{\varPhi}^\top (\mathbf{Y}_m - \mathbf{t}).$$

Southous tradition to Science and Technology 30/41

Due to the Laplace approximation

$$\int f(\mathbf{z})d\mathbf{z} \approx f(\mathbf{z}_0) \int \exp\left\{-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A}(\mathbf{z} - \mathbf{z}_0)\right\} d\mathbf{z} = f(\mathbf{z}_0) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}},$$

where

$$\mathbf{z}_0 = \arg\max_{\mathbf{z}} f(\mathbf{z}), \ \mathbf{A} = -\nabla\nabla \log f(\mathbf{z})\Big|_{\mathbf{z} = \mathbf{z}_0}$$

We get that

$$\nabla \log p(\mathbf{w}|\mathbf{Y}_m, \boldsymbol{\alpha}) = \boldsymbol{\Phi}^{\top}(\mathbf{Y}_m - \mathbf{t}) - \boldsymbol{\alpha}\mathbf{w},$$
$$\nabla \nabla \log p(\mathbf{w}|\mathbf{Y}_m, \boldsymbol{\alpha}) = -(\boldsymbol{\Phi}^{\top}\mathbf{R}\boldsymbol{\Phi} + \boldsymbol{\alpha}),$$

where ${\bf R}$ is an $m \times m$ diagonal matrix with elements $R_{ii} = t_i(1-t_i)$

We get that MAP estimate fulfils equality

$$\mathbf{\Phi}^{\top}(\mathbf{Y}_m - \mathbf{t}) - \mathbf{A}\mathbf{w} = 0 =$$

ullet We can use an approximation to get an estimate \mathbf{w}^*

$$\mathbf{w}^* = \mathbf{A}^{-1} \boldsymbol{\varPhi}^\top (\mathbf{Y}_m - \mathbf{t}).$$

4 D > 4 B > 4 B > 4 B > 9 Q ()

Skoltech

Due to the Laplace approximation

$$\int f(\mathbf{z})d\mathbf{z} \approx f(\mathbf{z}_0) \int \exp\left\{-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A}(\mathbf{z} - \mathbf{z}_0)\right\} d\mathbf{z} = f(\mathbf{z}_0) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}},$$

where

$$\mathbf{z}_0 = \arg\max_{\mathbf{z}} f(\mathbf{z}), \ \mathbf{A} = -\nabla\nabla \log f(\mathbf{z})\Big|_{\mathbf{z} = \mathbf{z}_0}$$

We get that

$$\nabla \log p(\mathbf{w}|\mathbf{Y}_m, \boldsymbol{\alpha}) = \boldsymbol{\Phi}^{\top}(\mathbf{Y}_m - \mathbf{t}) - \boldsymbol{\alpha}\mathbf{w},$$
$$\nabla \nabla \log p(\mathbf{w}|\mathbf{Y}_m, \boldsymbol{\alpha}) = -(\boldsymbol{\Phi}^{\top}\mathbf{R}\boldsymbol{\Phi} + \boldsymbol{\alpha}),$$

where ${f R}$ is an $m \times m$ diagonal matrix with elements $R_{ii} = t_i(1-t_i)$

We get that MAP estimate fulfils equality

$$\mathbf{\Phi}^{\top}(\mathbf{Y}_m - \mathbf{t}) - \mathbf{A}\mathbf{w} = 0 \Rightarrow$$

ullet We can use an approximation to get an estimate \mathbf{w}^*

$$\mathbf{w}^* = \mathbf{A}^{-1} \boldsymbol{\varPhi}^\top (\mathbf{Y}_m - \mathbf{t}).$$

Skoltech

30/41

RVM for Classification

Due to the Laplace approximation

$$\int f(\mathbf{z})d\mathbf{z} \approx f(\mathbf{z}_0) \int \exp\left\{-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^{\top} \mathbf{A}(\mathbf{z} - \mathbf{z}_0)\right\} d\mathbf{z} = f(\mathbf{z}_0) \frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}},$$

where

$$\mathbf{z}_0 = \arg\max_{\mathbf{z}} f(\mathbf{z}), \ \mathbf{A} = -\nabla\nabla \log f(\mathbf{z})\Big|_{\mathbf{z} = \mathbf{z}_0}$$

We get that

$$\nabla \log p(\mathbf{w}|\mathbf{Y}_m, \boldsymbol{\alpha}) = \boldsymbol{\Phi}^{\top}(\mathbf{Y}_m - \mathbf{t}) - \boldsymbol{\alpha}\mathbf{w},$$
$$\nabla \nabla \log p(\mathbf{w}|\mathbf{Y}_m, \boldsymbol{\alpha}) = -(\boldsymbol{\Phi}^{\top}\mathbf{R}\boldsymbol{\Phi} + \boldsymbol{\alpha}),$$

where ${f R}$ is an $m \times m$ diagonal matrix with elements $R_{ii} = t_i(1-t_i)$

We get that MAP estimate fulfils equality

$$\mathbf{\Phi}^{\top}(\mathbf{Y}_m - \mathbf{t}) - \mathbf{A}\mathbf{w} = 0 \Rightarrow$$

We can use an approximation to get an estimate w*:

$$\mathbf{w}^* = \mathbf{A}^{-1} \boldsymbol{\varPhi}^\top (\mathbf{Y}_m - \mathbf{t}).$$

one Institute of Science and Technology 30/41

• The maximum and the hessian of $\log p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\alpha)$ coincide with that of $\log p(\mathbf{w}|\mathbf{Y}_m, \boldsymbol{\alpha})$. Therefore, using the Laplace approximation we get that

$$-\frac{1}{2}(w_i^*)^2 + \frac{1}{2\alpha_i} - \frac{1}{2}[\mathbf{S}_m]_{ii} = 0$$

$$\alpha_i^{new} = \frac{\gamma_i}{(w_i^*)^2}$$

• For $\mathbf{z} = \boldsymbol{\Phi}\mathbf{w}^* - \mathbf{R}^{-1}(\mathbf{Y}_m - \mathbf{t})$ and $\mathbf{C} = \mathbf{R} + \boldsymbol{\Phi}\mathbf{A}\boldsymbol{\Phi}^{\top}$ we get

$$\log p(\mathbf{Y}_m | \boldsymbol{\alpha}) = -\frac{1}{2} \left\{ m \log(2\pi) + \log |\mathbf{C}| + (\mathbf{z}\mathbf{C}^{-1}\mathbf{z}) \right\}.$$

イロト イ部ト イミト イミト

31/41

• The maximum and the hessian of $\log p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})$ coincide with that of $\log p(\mathbf{w}|\mathbf{Y}_m,\boldsymbol{\alpha})$. Therefore, using the Laplace approximation we get that

$$p(\mathbf{Y}_m|\boldsymbol{\alpha}) = \int p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w}$$

Setting the derivative of the marginal likelihood to zero we get

$$-\frac{1}{2}(w_i^*)^2 + \frac{1}{2\alpha_i} - \frac{1}{2}[\mathbf{S}_m]_{ii} = 0$$

• For $\gamma_i = 1 - \alpha_i [\mathbf{S}_m]_{ii}$ we get

$$\alpha_i^{new} = \frac{\gamma_i}{(w_i^*)^2}$$

ullet For $\mathbf{z} = oldsymbol{arPhi} \mathbf{w}^* - \mathbf{R}^{-1} (\mathbf{Y}_m - \mathbf{t})$ and $\mathbf{C} = \mathbf{R} + oldsymbol{arPhi} \mathbf{A} oldsymbol{arPhi}^ op$ we get

$$\log p(\mathbf{Y}_m | \boldsymbol{\alpha}) = -\frac{1}{2} \left\{ m \log(2\pi) + \log |\mathbf{C}| + (\mathbf{z}\mathbf{C}^{-1}\mathbf{z}) \right\}.$$

31/41

The same form as in the regression case ⇒ we can apply the same analysis of sparsity and obtain the same fast learning algorithm

We iterate between twing of we and on the same fast learning algorithm

• The maximum and the hessian of $\log p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})$ coincide with that of $\log p(\mathbf{w}|\mathbf{Y}_m,\boldsymbol{\alpha})$. Therefore, using the Laplace approximation we get that

$$p(\mathbf{Y}_m|\boldsymbol{\alpha}) = \int p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \approx p(\mathbf{Y}_m|\mathbf{w}^*)p(\mathbf{w}^*|\boldsymbol{\alpha})(2\pi)^{M/2}|\mathbf{S}_m|^{\frac{1}{2}}$$

Setting the derivative of the marginal likelihood to zero we get

$$-\frac{1}{2}(w_i^*)^2 + \frac{1}{2\alpha_i} - \frac{1}{2}[\mathbf{S}_m]_{ii} = 0$$

• For $\gamma_i = 1 - \alpha_i [\mathbf{S}_m]_{ii}$ we get

$$\alpha_i^{new} = \frac{\gamma_i}{(w_i^*)^2}$$

 \bullet For $\mathbf{z} = \mathbf{\varPhi}\mathbf{w}^* - \mathbf{R}^{-1}(\mathbf{Y}_m - \mathbf{t})$ and $\mathbf{C} = \mathbf{R} + \mathbf{\varPhi}\mathbf{A}\mathbf{\varPhi}^\top$ we get

$$\log p(\mathbf{Y}_m | \boldsymbol{\alpha}) = -\frac{1}{2} \left\{ m \log(2\pi) + \log |\mathbf{C}| + (\mathbf{z}\mathbf{C}^{-1}\mathbf{z}) \right\}.$$

31/41

The same form as in the regression case ⇒ we can apply the same analysis of sparsity and obtain the same fast learning algorithm

We iterate between twing of we and on the same fast learning algorithm

• The maximum and the hessian of $\log p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})$ coincide with that of $\log p(\mathbf{w}|\mathbf{Y}_m,\boldsymbol{\alpha})$. Therefore, using the Laplace approximation we get that

$$p(\mathbf{Y}_m | \boldsymbol{\alpha}) = \int p(\mathbf{Y}_m | \mathbf{w}) p(\mathbf{w} | \boldsymbol{\alpha}) d\mathbf{w} \approx p(\mathbf{Y}_m | \mathbf{w}^*) p(\mathbf{w}^* | \boldsymbol{\alpha}) (2\pi)^{M/2} |\mathbf{S}_m|^{\frac{1}{2}}$$
$$\mathbf{w}^* = \boldsymbol{\alpha}^{-1} \boldsymbol{\Phi}^\top (\mathbf{Y}_m - \mathbf{t}),$$

Setting the derivative of the marginal likelihood to zero we get

$$-\frac{1}{2}(w_i^*)^2 + \frac{1}{2\alpha_i} - \frac{1}{2}[\mathbf{S}_m]_{ii} = 0$$

ullet For $\gamma_i=1-lpha_i[{f S}_m]_{ii}$ we get

$$\alpha_i^{new} = \frac{\gamma_i}{(w_i^*)^2}$$

 \bullet For $\mathbf{z} = \mathbf{\varPhi}\mathbf{w}^* - \mathbf{R}^{-1}(\mathbf{Y}_m - \mathbf{t})$ and $\mathbf{C} = \mathbf{R} + \mathbf{\varPhi}\mathbf{A}\mathbf{\varPhi}^\top$ we get

$$\log p(\mathbf{Y}_m | \boldsymbol{\alpha}) = -\frac{1}{2} \left\{ m \log(2\pi) + \log |\mathbf{C}| + (\mathbf{z}\mathbf{C}^{-1}\mathbf{z}) \right\}.$$

The same form as in the regression case \Rightarrow we can apply the same analysis of sparsity and obtain the same fast learning algorithm

We iterate between tuning of we and on the same fast learning algorithm

31/41

• The maximum and the hessian of $\log p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})$ coincide with that of $\log p(\mathbf{w}|\mathbf{Y}_m,\boldsymbol{\alpha})$. Therefore, using the Laplace approximation we get that

$$p(\mathbf{Y}_m|\boldsymbol{\alpha}) = \int p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \approx p(\mathbf{Y}_m|\mathbf{w}^*)p(\mathbf{w}^*|\boldsymbol{\alpha})(2\pi)^{M/2}|\mathbf{S}_m|^{\frac{1}{2}}$$
$$\mathbf{w}^* = \boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^\top(\mathbf{Y}_m - \mathbf{t}), \ \mathbf{S}_m = (\boldsymbol{\Phi}^\top\mathbf{R}\boldsymbol{\Phi} + \boldsymbol{\alpha})^{-1}$$

Setting the derivative of the marginal likelihood to zero we get

$$-\frac{1}{2}(w_i^*)^2 + \frac{1}{2\alpha_i} - \frac{1}{2}[\mathbf{S}_m]_{ii} = 0$$

• For $\gamma_i = 1 - \alpha_i [\mathbf{S}_m]_{ii}$ we get

$$\alpha_i^{new} = \frac{\gamma_i}{(w_i^*)^2}$$

 $\bullet \ \, \mathsf{For} \,\, \mathbf{z} = \mathbf{\varPhi} \mathbf{w}^* - \mathbf{R}^{-1} (\mathbf{Y}_m - \mathbf{t}) \,\, \mathsf{and} \,\, \mathbf{C} = \mathbf{R} + \mathbf{\varPhi} \mathbf{A} \mathbf{\varPhi}^\top \,\, \mathsf{we} \,\, \mathsf{get}$

$$\log p(\mathbf{Y}_m | \boldsymbol{\alpha}) = -\frac{1}{2} \left\{ m \log(2\pi) + \log |\mathbf{C}| + (\mathbf{z}\mathbf{C}^{-1}\mathbf{z}) \right\}.$$

The same form as in the regression case \Rightarrow we can apply the same analysis of sparsity and obtain the same fast learning algorithm

We iterate between tuning of we and on the same fast learning algorithm

31/41

Burnaev, Bayesian ML

Skoltech

• The maximum and the hessian of $\log p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})$ coincide with that of $\log p(\mathbf{w}|\mathbf{Y}_m,\boldsymbol{\alpha})$. Therefore, using the Laplace approximation we get that

$$p(\mathbf{Y}_m|\boldsymbol{\alpha}) = \int p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \approx p(\mathbf{Y}_m|\mathbf{w}^*)p(\mathbf{w}^*|\boldsymbol{\alpha})(2\pi)^{M/2}|\mathbf{S}_m|^{\frac{1}{2}}$$
$$\mathbf{w}^* = \boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^{\top}(\mathbf{Y}_m - \mathbf{t}), \ \mathbf{S}_m = (\boldsymbol{\Phi}^{\top}\mathbf{R}\boldsymbol{\Phi} + \boldsymbol{\alpha})^{-1}$$

Setting the derivative of the marginal likelihood to zero we get

$$-\frac{1}{2}(w_i^*)^2 + \frac{1}{2\alpha_i} - \frac{1}{2}[\mathbf{S}_m]_{ii} = 0$$

• For $\gamma_i = 1 - \alpha_i [\mathbf{S}_m]_{ii}$ we get

$$\alpha_i^{new} = \frac{\gamma_i}{(w_i^*)^2}$$

 \bullet For $\mathbf{z}=\mathbf{\varPhi}\mathbf{w}^*-\mathbf{R}^{-1}(\mathbf{Y}_m-\mathbf{t})$ and $\mathbf{C}=\mathbf{R}+\mathbf{\varPhi}\mathbf{A}\mathbf{\varPhi}^\top$ we get

$$\log p(\mathbf{Y}_m | \boldsymbol{\alpha}) = -\frac{1}{2} \left\{ m \log(2\pi) + \log |\mathbf{C}| + (\mathbf{z}\mathbf{C}^{-1}\mathbf{z}) \right\}.$$

31/41

The same form as in the regression case \Rightarrow we can apply the same analysis of sparsity and obtain the same fast learning algorithm

Burnaev, Bayesian ML Skoltech

• The maximum and the hessian of $\log p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})$ coincide with that of $\log p(\mathbf{w}|\mathbf{Y}_m,\boldsymbol{\alpha})$. Therefore, using the Laplace approximation we get that

$$p(\mathbf{Y}_m|\boldsymbol{\alpha}) = \int p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \approx p(\mathbf{Y}_m|\mathbf{w}^*)p(\mathbf{w}^*|\boldsymbol{\alpha})(2\pi)^{M/2}|\mathbf{S}_m|^{\frac{1}{2}}$$

$$\mathbf{w}^* = \boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^{\top}(\mathbf{Y}_m - \mathbf{t}), \, \mathbf{S}_m = (\boldsymbol{\Phi}^{\top}\mathbf{R}\boldsymbol{\Phi} + \boldsymbol{\alpha})^{-1}$$

Setting the derivative of the marginal likelihood to zero we get

$$-\frac{1}{2}(w_i^*)^2 + \frac{1}{2\alpha_i} - \frac{1}{2}[\mathbf{S}_m]_{ii} = 0$$

• For $\gamma_i = 1 - \alpha_i [\mathbf{S}_m]_{ii}$ we get

$$\alpha_i^{new} = \frac{\gamma_i}{(w_i^*)^2}$$

 $\bullet \ \, \mathsf{For} \,\, \mathbf{z} = \mathbf{\varPhi} \mathbf{w}^* - \mathbf{R}^{-1} (\mathbf{Y}_m - \mathbf{t}) \,\, \mathsf{and} \,\, \mathbf{C} = \mathbf{R} + \mathbf{\varPhi} \mathbf{A} \mathbf{\varPhi}^\top \,\, \mathsf{we} \,\, \mathsf{get}$

$$\log p(\mathbf{Y}_m | \boldsymbol{\alpha}) = -\frac{1}{2} \left\{ m \log(2\pi) + \log |\mathbf{C}| + (\mathbf{z}\mathbf{C}^{-1}\mathbf{z}) \right\}.$$

31/41

The same form as in the regression case \Rightarrow we can apply the same analysis of sparsity and obtain the same fast learning algorithm

• The maximum and the hessian of $\log p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})$ coincide with that of $\log p(\mathbf{w}|\mathbf{Y}_m,\boldsymbol{\alpha})$. Therefore, using the Laplace approximation we get that

$$p(\mathbf{Y}_m|\boldsymbol{\alpha}) = \int p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \approx p(\mathbf{Y}_m|\mathbf{w}^*)p(\mathbf{w}^*|\boldsymbol{\alpha})(2\pi)^{M/2}|\mathbf{S}_m|^{\frac{1}{2}}$$
$$\mathbf{w}^* = \boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^\top(\mathbf{Y}_m - \mathbf{t}), \ \mathbf{S}_m = (\boldsymbol{\Phi}^\top\mathbf{R}\boldsymbol{\Phi} + \boldsymbol{\alpha})^{-1}$$

Setting the derivative of the marginal likelihood to zero we get

$$-\frac{1}{2}(w_i^*)^2 + \frac{1}{2\alpha_i} - \frac{1}{2}[\mathbf{S}_m]_{ii} = 0$$

• For $\gamma_i = 1 - \alpha_i [\mathbf{S}_m]_{ii}$ we get

$$\alpha_i^{new} = \frac{\gamma_i}{(w_i^*)^2}$$

 \bullet For $\mathbf{z} = \mathbf{\varPhi}\mathbf{w}^* - \mathbf{R}^{-1}(\mathbf{Y}_m - \mathbf{t})$ and $\mathbf{C} = \mathbf{R} + \mathbf{\varPhi}\mathbf{A}\mathbf{\varPhi}^\top$ we get

$$\log p(\mathbf{Y}_m | \boldsymbol{\alpha}) = -\frac{1}{2} \left\{ m \log(2\pi) + \log |\mathbf{C}| + (\mathbf{z} \mathbf{C}^{-1} \mathbf{z}) \right\}.$$

The same form as in the regression case \Rightarrow we can apply the same analysis of sparsity and obtain the same fast learning algorithm

Burnaev, Bayesian ML

• The maximum and the hessian of $\log p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})$ coincide with that of $\log p(\mathbf{w}|\mathbf{Y}_m,\boldsymbol{\alpha})$. Therefore, using the Laplace approximation we get that

$$p(\mathbf{Y}_m|\boldsymbol{\alpha}) = \int p(\mathbf{Y}_m|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} \approx p(\mathbf{Y}_m|\mathbf{w}^*)p(\mathbf{w}^*|\boldsymbol{\alpha})(2\pi)^{M/2}|\mathbf{S}_m|^{\frac{1}{2}}$$
$$\mathbf{w}^* = \boldsymbol{\alpha}^{-1}\boldsymbol{\varPhi}^\top(\mathbf{Y}_m - \mathbf{t}), \ \mathbf{S}_m = (\boldsymbol{\varPhi}^\top\mathbf{R}\boldsymbol{\varPhi} + \boldsymbol{\alpha})^{-1}$$

Setting the derivative of the marginal likelihood to zero we get

$$-\frac{1}{2}(w_i^*)^2 + \frac{1}{2\alpha_i} - \frac{1}{2}[\mathbf{S}_m]_{ii} = 0$$

• For $\gamma_i = 1 - \alpha_i [\mathbf{S}_m]_{ii}$ we get

$$\alpha_i^{new} = \frac{\gamma_i}{(w_i^*)^2}$$

 \bullet For $\mathbf{z} = \mathbf{\varPhi}\mathbf{w}^* - \mathbf{R}^{-1}(\mathbf{Y}_m - \mathbf{t})$ and $\mathbf{C} = \mathbf{R} + \mathbf{\varPhi}\mathbf{A}\mathbf{\varPhi}^\top$ we get

$$\log p(\mathbf{Y}_m | \boldsymbol{\alpha}) = -\frac{1}{2} \left\{ m \log(2\pi) + \log |\mathbf{C}| + (\mathbf{z}\mathbf{C}^{-1}\mathbf{z}) \right\}.$$

The same form as in the regression case \Rightarrow we can apply the same analysis of sparsity and obtain the same fast learning algorithm

ullet We iterate between tuning of ${f w}$ and ${f lpha}$

Skoltech

K-class classification case

• We set $a_k = \mathbf{w}_k \cdot \mathbf{x}^{\top}$ and probabilities of specific classes to be equal to

$$t_k(\mathbf{x}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

The likelihood function is then given by

$$\log p(\mathbf{Y}_m|\mathbf{w}_1,\ldots,\mathbf{w}_K) = \prod_{i=1}^m \prod_{k=1}^K t_{ik}^{y_{ik}}$$

All other steps are exactly the same

Skoltech

- K-class classification case
- We set $a_k = \mathbf{w}_k \cdot \mathbf{x}^{\top}$ and probabilities of specific classes to be equal to

$$t_k(\mathbf{x}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

The likelihood function is then given by

$$\log p(\mathbf{Y}_m|\mathbf{w}_1,\ldots,\mathbf{w}_K) = \prod_{i=1}^m \prod_{k=1}^K t_{ik}^{y_{ik}}$$

All other steps are exactly the same!



SKOITECH Bulloon trettare of Science and Technology 32/41

- K-class classification case
- We set $a_k = \mathbf{w}_k \cdot \mathbf{x}^{\top}$ and probabilities of specific classes to be equal to

$$t_k(\mathbf{x}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

The likelihood function is then given by

$$\log p(\mathbf{Y}_m | \mathbf{w}_1, \dots, \mathbf{w}_K) = \prod_{i=1}^m \prod_{k=1}^K t_{ik}^{y_{ik}}$$

All other steps are exactly the same

32/41

- K-class classification case
- We set $a_k = \mathbf{w}_k \cdot \mathbf{x}^{\top}$ and probabilities of specific classes to be equal to

$$t_k(\mathbf{x}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

The likelihood function is then given by

$$\log p(\mathbf{Y}_m | \mathbf{w}_1, \dots, \mathbf{w}_K) = \prod_{i=1}^m \prod_{k=1}^K t_{ik}^{y_{ik}}$$

All other steps are exactly the same!

- Bayesian Linear Models for Classification
- 2 Laplace Approximation
- 3 Bayesian Logistic Regression
- Relevance Vector Machine for Classification
- 5 RVM application examples

- A state-of-the-art method for classification and regression
- ullet Given data set comprising m input vectors ${f x}_i$, model has the form

$$f(\mathbf{x}; \mathbf{w}) = \sum_{i=1}^{m} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0$$

- As many kernel functions $K(\cdot, \mathbf{x}_i)$ as examples, i.e. M = m + 1 parameters plus kernel width
- Support vector learning: minimize objective function of the form

$$E(\mathbf{w}) = E_D(\mathbf{w}) - \lambda \times \text{(size of margin)}$$

- gives excellent accuracy (particular in classification)
- as a side-effect, many \mathbf{w}_i get set to zero the model is sparse
- RVM is simply a Bayesian model utilising the same data dependent kernel basis as the SVM

$$t(\boldsymbol{\phi}) = \sum_{i=1}^{m} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0$$

- A state-of-the-art method for classification and regression
- ullet Given data set comprising m input vectors ${f x}_i$, model has the form

$$f(\mathbf{x}; \mathbf{w}) = \sum_{i=1}^{m} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0$$

- As many kernel functions $K(\cdot, \mathbf{x}_i)$ as examples, i.e. M=m+1 parameters plus kernel width
- Support vector learning: minimize objective function of the form

$$E(\mathbf{w}) = E_D(\mathbf{w}) - \lambda \times \text{(size of margin)}$$

- gives excellent accuracy (particular in classification)
- as a side-effect, many \mathbf{w}_i get set to zero the model is sparse
- RVM is simply a Bayesian model utilising the same data dependent kernel basis as the SVM

$$t(\phi) = \sum_{i=1}^{m} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0$$

- A state-of-the-art method for classification and regression
- ullet Given data set comprising m input vectors ${f x}_i$, model has the form

$$f(\mathbf{x}; \mathbf{w}) = \sum_{i=1}^{m} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0$$

- As many kernel functions $K(\cdot, \mathbf{x}_i)$ as examples, i.e. M=m+1 parameters plus kernel width
- Support vector learning: minimize objective function of the form

$$E(\mathbf{w}) = E_D(\mathbf{w}) - \lambda \times \text{(size of margin)}$$

- gives excellent accuracy (particular in classification)
- as a side-effect, many \mathbf{w}_i get set to zero the model is sparse
- RVM is simply a Bayesian model utilising the same data dependent kernel basis as the SVM

$$t(\phi) = \sum_{i=1}^{m} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0$$

- A state-of-the-art method for classification and regression
- ullet Given data set comprising m input vectors ${f x}_i$, model has the form

$$f(\mathbf{x}; \mathbf{w}) = \sum_{i=1}^{m} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0$$

- \bullet As many kernel functions $K(\cdot,\mathbf{x}_i)$ as examples, i.e. M=m+1 parameters plus kernel width
- Support vector learning: minimize objective function of the form

$$E(\mathbf{w}) = E_D(\mathbf{w}) - \lambda \times \text{(size of margin)}$$

- gives excellent accuracy (particular in classification)
- as a side-effect, many \mathbf{w}_i get set to zero the model is sparse
- RVM is simply a Bayesian model utilising the same data dependent kernel basis as the SVM

$$t(\phi) = \sum_{i=1}^{m} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0$$

Burnaev, Bayesian ML Guidan totra efforce safetheig

- A state-of-the-art method for classification and regression
- ullet Given data set comprising m input vectors ${f x}_i$, model has the form

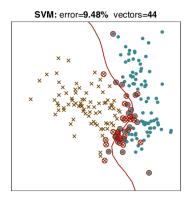
$$f(\mathbf{x}; \mathbf{w}) = \sum_{i=1}^{m} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0$$

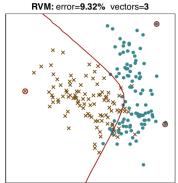
- \bullet As many kernel functions $K(\cdot,\mathbf{x}_i)$ as examples, i.e. M=m+1 parameters plus kernel width
- Support vector learning: minimize objective function of the form

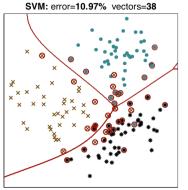
$$E(\mathbf{w}) = E_D(\mathbf{w}) - \lambda \times (\text{size of margin})$$

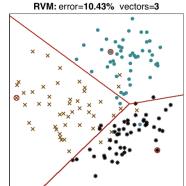
- gives excellent accuracy (particular in classification)
- as a side-effect, many \mathbf{w}_i get set to zero the model is sparse
- RVM is simply a Bayesian model utilising the same data dependent kernel basis as the SVM

$$t(\boldsymbol{\phi}) = \sum_{i=1}^{m} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0$$









Classification Performance Illustration

			errors		$_$ vectors $_$	
Data set	N	d	SVM	RVM	SVM	RVM
Pima Diabetes	200	8	20.1%	19.6%	109	4
U.S.P.S.	7291	256	4.4%	5.1%	2540	316
Banana	400	2	10.9%	10.8%	135.2	11.4
Breast Cancer	200	9	26.9%	29.9%	116.7	6.3
Titanic	150	3	22.1%	23.0%	93.7	65.3
Waveform	400	21	10.3%	10.9%	146.4	14.6
German	700	20	22.6%	22.2%	411.2	12.5
Image	1300	18	3.0%	3.9 %	166.6	34.6
Normalised Mean			1.00	1.08	1.00	0.17

Comparison with the SVM

- General observations:
 - RVM gives better generalization in regression (?)
 - RVM gives better generalization in classification (?)
 - RVM is much sparse (but the SVM is not designed to be sparse)
- There are other advantages of a Bayesian approach:
 - no "nuisance" parameters to set
 - posterior probabilities in classification
 - error bars in regression
 - principled method for more than two classes
 - not limited to Mercer kernels
 - potential to estimate input scale parameters and compare kernels

Comparison with the SVM

- General observations:
 - RVM gives better generalization in regression (?)
 - RVM gives better generalization in classification (?)
 - RVM is much sparse (but the SVM is not designed to be sparse)
- There are other advantages of a Bayesian approach:
 - no "nuisance" parameters to set
 - posterior probabilities in classification
 - error bars in regression
 - principled method for more than two classes
 - not limited to Mercer kernels
 - potential to estimate input scale parameters and compare kernels

- As in the SVM, we must choose the kernel and set any associated parameters
- A Bayesian could compare alternative kernels by computing the fully marginalized probabilities of the data under candidate models, e.g.

$$p(\mathbf{Y}_m|K_1) = \int p(\mathbf{Y}_m|\boldsymbol{\alpha}, K_1)p(\boldsymbol{\alpha}|K_1)d\boldsymbol{\alpha}$$

- We already know this integral isn't analytically tractable
- ullet Approximation via sampling is not feasible for multiple lpha
- Deterministic approximations to this integral have proved inaccurate
- ullet But $p(\mathbf{Y}_m|oldsymbol{lpha}_{MAP},K)$ is a "reasonable" criterion for choosing kernels

↓□ > ∢部 > ∢差 > く差 > を か を のへで

Burnaev, Bayesian ML Skoltech

Choosing the kernel

- As in the SVM, we must choose the kernel and set any associated parameters
- A Bayesian could compare alternative kernels by computing the fully marginalized probabilities of the data under candidate models, e.g.

$$p(\mathbf{Y}_m|K_1) = \int p(\mathbf{Y}_m|\boldsymbol{\alpha}, K_1)p(\boldsymbol{\alpha}|K_1)d\boldsymbol{\alpha}$$

- We already know this integral isn't analytically tractable
- ullet Approximation via sampling is not feasible for multiple lpha
- Deterministic approximations to this integral have proved inaccurate
- ullet But $p(\mathbf{Y}_m | \boldsymbol{\alpha}_{MAP}, K)$ is a "reasonable" criterion for choosing kernels

4□ > 4回 > 4 = > 4 = > = の4 ○

Skoltech

- As in the SVM, we must choose the kernel and set any associated parameters
- A Bayesian could compare alternative kernels by computing the fully marginalized probabilities of the data under candidate models, e.g.

$$p(\mathbf{Y}_m|K_1) = \int p(\mathbf{Y}_m|\boldsymbol{\alpha}, K_1)p(\boldsymbol{\alpha}|K_1)d\boldsymbol{\alpha}$$

- We already know this integral isn't analytically tractable
- ullet Approximation via sampling is not feasible for multiple lpha
- Deterministic approximations to this integral have proved inaccurate
- ullet But $p(\mathbf{Y}_m | \boldsymbol{\alpha}_{MAP}, K)$ is a "reasonable" criterion for choosing kernels

4□ > 4回 > 4 回 > 4 回 > 1 回 の Q ○

Skoltech

- As in the SVM, we must choose the kernel and set any associated parameters
- A Bayesian could compare alternative kernels by computing the fully marginalized probabilities of the data under candidate models, e.g.

$$p(\mathbf{Y}_m|K_1) = \int p(\mathbf{Y}_m|\boldsymbol{\alpha}, K_1)p(\boldsymbol{\alpha}|K_1)d\boldsymbol{\alpha}$$

- We already know this integral isn't analytically tractable
- ullet Approximation via sampling is not feasible for multiple lpha
- Deterministic approximations to this integral have proved inaccurate
- ullet But $p(\mathbf{Y}_m | \boldsymbol{\alpha}_{MAP}, K)$ is a "reasonable" criterion for choosing kernels

4□ > 4回 > 4 回 > 4 回 > 1 回 の Q ○

Burnaev, Bayesian ML Skottech

- As in the SVM, we must choose the kernel and set any associated parameters
- A Bayesian could compare alternative kernels by computing the fully marginalized probabilities of the data under candidate models, e.g.

$$p(\mathbf{Y}_m|K_1) = \int p(\mathbf{Y}_m|\boldsymbol{\alpha}, K_1)p(\boldsymbol{\alpha}|K_1)d\boldsymbol{\alpha}$$

- We already know this integral isn't analytically tractable
- ullet Approximation via sampling is not feasible for multiple lpha
- Deterministic approximations to this integral have proved inaccurate
- But $p(\mathbf{Y}_m | \boldsymbol{\alpha}_{MAP}, K)$ is a "reasonable" criterion for choosing kernels

Burnaev, Bayesian ML Skoltech

- As in the SVM, we must choose the kernel and set any associated parameters
- A Bayesian could compare alternative kernels by computing the fully marginalized probabilities of the data under candidate models, e.g.

$$p(\mathbf{Y}_m|K_1) = \int p(\mathbf{Y}_m|\boldsymbol{\alpha}, K_1)p(\boldsymbol{\alpha}|K_1)d\boldsymbol{\alpha}$$

- We already know this integral isn't analytically tractable
- ullet Approximation via sampling is not feasible for multiple lpha
- Deterministic approximations to this integral have proved inaccurate
- ullet But $p(\mathbf{Y}_m|oldsymbol{lpha}_{MAP},K)$ is a "reasonable" criterion for choosing kernels

4□ > 4回 > 4 = > 4 = > = 900

39/41

Burnaev, Bayesian ML

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\{-\eta \|\mathbf{x} - \mathbf{x}_i\|^2\}$$

- lacksquare SVM: cross-validation can be used to set scale parameter η
- RVM: we can optimize the marginal likelihood function w.r.t. η
- Furthermore we can optimize multiple scale parameters η , one for each of the d input dimensions

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\left\{-\sum_{k=1}^d \eta_k (\mathbf{x}_k - \mathbf{x}_{ik})^2\right\}$$

 Implementing sparsity of input variables (q.v. Gaussian process models)

ロト (個) (注) (注) 注 の(の

rotage of Science and Technology 40 / 41

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\{-\eta \|\mathbf{x} - \mathbf{x}_i\|^2\}$$

- ullet SVM: cross-validation can be used to set scale parameter η
- RVM: we can optimize the marginal likelihood function w.r.t. η
- ullet Furthermore we can optimize multiple scale parameters η , one for each of the d input dimensions

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\left\{-\sum_{k=1}^d \eta_k (\mathbf{x}_k - \mathbf{x}_{ik})^2\right\}$$

 Implementing sparsity of input variables (q.v. Gaussian process models)

Burnaev, Bayesian ML Subsection of the control of t

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\{-\eta \|\mathbf{x} - \mathbf{x}_i\|^2\}$$

- ullet SVM: cross-validation can be used to set scale parameter η
- ullet RVM: we can optimize the marginal likelihood function w.r.t. η
- Furthermore we can optimize multiple scale parameters η , one for each of the d input dimensions

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\left\{-\sum_{k=1}^d \eta_k (\mathbf{x}_k - \mathbf{x}_{ik})^2\right\}$$

 Implementing sparsity of input variables (q.v. Gaussian process models)

4 D > 4 B > 4 B > B 9 9 9

Burnaev, Bayesian ML Skottech Microscotticology 4

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\{-\eta \|\mathbf{x} - \mathbf{x}_i\|^2\}$$

- ullet SVM: cross-validation can be used to set scale parameter η
- ullet RVM: we can optimize the marginal likelihood function w.r.t. η
- ullet Furthermore we can optimize multiple scale parameters η , one for each of the d input dimensions

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\left\{-\sum_{k=1}^d \eta_k (\mathbf{x}_k - \mathbf{x}_{ik})^2\right\}$$

 Implementing sparsity of input variables (q.v. Gaussian process models)

40/41

Burnaev, Bayesian ML

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\{-\eta \|\mathbf{x} - \mathbf{x}_i\|^2\}$$

- ullet SVM: cross-validation can be used to set scale parameter η
- ullet RVM: we can optimize the marginal likelihood function w.r.t. η
- ullet Furthermore we can optimize multiple scale parameters η , one for each of the d input dimensions

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\left\{-\sum_{k=1}^d \eta_k (\mathbf{x}_k - \mathbf{x}_{ik})^2\right\}$$

 Implementing sparsity of input variables (q.v. Gaussian process models)

(□▶ ◀♬▶ ◀불▶ ◀불▶ 불 쒸٩♡

Burnaev, Bayesian ML

Impact on Regression Benchmarks

ullet η -RVR: optimization over both lpha and η

	Test error			# kernels		
Dataset	SVR	RVR	η -RVR	SVR	RVR	η -RVR
Friedman #1	2.92	2.80	0.27	116.6	59.4	11.5
Friedman #2	4140	3505	2593	110.3	6.9	3.9
Friedman #3	0.0202	0.0164	0.0119	106.5	11.5	6.4

• Friedman No. 1: 10-dimensional input space, but functions depends only on variables 1-5. Final η -values shown below

Impact on Regression Benchmarks

ullet η -RVR: optimization over both lpha and η

	Test error			# kernels		
Dataset	SVR	RVR	η -RVR	SVR	RVR	η -RVR
Friedman #1	2.92	2.80	0.27	116.6	59.4	11.5
Friedman #2	4140	3505	2593	110.3	6.9	3.9
Friedman #3	0.0202	0.0164	0.0119	106.5	11.5	6.4

• Friedman No. 1: 10-dimensional input space, but functions depends only on variables 1-5. Final η -values shown below

