UNIVERSITY OF AMSTERDAM

Faculty of Science

Exam

Machine Learning 1

Midterm Exam
Date: September 29, 2017
Time: 13:00-15:00

Number of pages: 9 (including front page)
Number of questions: 5
Maximum number of points to earn: 41
At each question the number of points you can earn is indicated.

BEFORE YOU START

- Please **wait** until you are instructed to open the booklet.

- Check if your version of the exam is complete.

- Write down **your name, student ID number**, and if applicable the **version number** on **each sheet** that you hand in. Also **number the pages**.

- Your **mobile phone** has to be switched off and in the coat or bag. Your **coat and bag** must be under your table.

- **Tools allowed**: 1 handwritten double-sided A4-size cheat sheet, pen.

PRACTICAL MATTERS

- The first 30 minutes and the last 15 minutes you are not allowed to leave the room, not even to visit the toilet.

- You are obliged to identify yourself at the request of the examiner (or his representative) with a proof of your enrollment or a valid ID.

- During the examination it is not permitted to visit the toilet, unless the proctor gives permission to do so.

- 15 minutes before the end, you will be warned that the time to hand in is approaching.

- If applicable, please fill out the evaluation form at the end of the exam.

Good luck!

1 Multiple Choice Questions

/16

For the evaluation of each question note: several answers might be correct and at least one is correct. You are granted one point if every correct answer is 'marked' **and** every incorrect answer is 'not marked'. For each mistake a 1/2 point is deducted, with the minimum possible number of points per question equal to 0. A box counts as 'marked' if a clearly visible symbol is written in there or if the box is blackened out. In the case you want to change an already marked box write 'not marked' next to the box.

1. Which of the following tasks is a supervised learning task? **/1**

   ☐ Grouping users of a movie streaming service with a clustering algorithm.

   ☒ Predicting the house prices in Amsterdam based on data obtained from the last 10 years.

   ☒ Classifying birds in natural images.

   ☐ Discovering different types of spam emails in a large collection of spam emails.

2. The Gaussian multivariate distribution for a random variable $\mathbf{x} \in \mathbb{R}^D$ is given by: $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$. Which of the following statements are correct? **/1**

   ☐ $\boldsymbol{\Sigma}$ is the covariance matrix of the multivariate Gaussian distribution and is equal to $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$, with $\mathbf{x}$ sampled from the multivariate Gaussian distribution.

   ☒ If $\mathbf{x}$ is a vector of size D (i.e. $\mathbf{x} \in \mathbb{R}^D$), then $\boldsymbol{\mu}$ is also a vector of size D.

   ☒ Let us denote $\mathbf{x} = (x_1, ..., x_D)^T$. If $\boldsymbol{\Sigma}$ is a matrix with nonzero entries on its diagonal, and zero-valued entries for all off-diagonal elements, then $\text{cov}[x_i, x_j] = 0$ for $i \neq j$.

3. Which of the following expressions are correct, given no independence assumption and for (non-trivial) discrete random variables? **/1**

   ☐ $\sum_b P(A|B = b) = 1$.

   ☒ For two values $a_1 \neq a_2$, $P(A = a_1 \text{ or } A = a_2|B) = P(A = a_1|B) + P(A = a_2|B)$.

   ☒ $\sum_a \sum_b P(A = a|B = b)P(B = b) = 1$.

   ☐ None of the above.

4. Which of the following equations are correct? **/1**

   ☐ $p(x) = \int p(x|y)dy$.

   ☒ The probability that a continuous random variable x takes on a value on the interval (a, b) with $b > a$ is given by $p(x \in (a, b)) = \int_a^b p(x)dx$.

   ☐ $p(x, y) = p(x|y)p(x)$.

   ☐ None of the above.

5. Let $\sigma(a) = \frac{1}{1+e^{-a}}$ be the sigmoid function. Which of the following statements about the sigmoid function are correct? **/1**

   [X] $\frac{d\sigma(a)}{da} = \sigma(a)(1 - \sigma(a))$.

   [ ] $\sigma(-a) = -\sigma(a)$.

   [ ] $\frac{\sigma(a)-1}{\sigma(a)} = \sigma(a)(\sigma(a) - 1)$.

   [X] $\sigma(-a) = 1 - \sigma(a)$.

6. You are given a dataset $\{\boldsymbol{x}_n, t_n\}$ of N datapoints, and basis-functions $\boldsymbol{\phi} = (\phi_0, \dots, \phi_{M-1})^T$, and you will perform regularized linear regression with the error function $E(\mathbf{w}, \lambda) = \frac{1}{2} \sum_{n=1}^{N} \left(\boldsymbol{\phi}(\mathbf{x}_n)^T \mathbf{w} - t_n \right)^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}$. You will use K-fold cross validation to determine the best value of the regularization parameter λ. Suppose λ is allowed to take on the values $\lambda_1, \lambda_2, \lambda_3$ or λ_4. Which of the following statements are correct? **/1**

   [X] For K-fold cross validation you split your datapoints into K different folds.

   [ ] For each possible value of λ, you will have to train your model with that value of λ on $K - 1$ different splits of your dataset.

   [X] In order to determine the best value of λ, you have to train your regularized linear regression model a total number of $4K$ times.

   [X] If $K = N$ this type of cross validation is also called leave-one-out cross validation.

   [ ] Once you have determined the best value of λ using K-fold cross validation on the dataset $\{\boldsymbol{x}_n, t_n\}$, you should use the average cross validation score that you obtained for the best λ to report the performance of your algorithm.

7. Consider again regularized linear regression with the error function
   $E(\mathbf{w}, \lambda) = \frac{1}{2N} \sum_{n=1}^{N} \left(\boldsymbol{\phi}(\mathbf{x}_n)^T \mathbf{w} - t_n \right)^2 + \frac{\lambda}{2} \mathbf{w}^T \mathbf{w}$ (Note the $1/N$ factor). You want to find the optimal regularization penalty $\lambda \in \{1, 0.1, 0.01\}$, but this time you will split your data into a training set, a validation set and a test set. You obtain the following validation and training errors $E_{\text{val}}(\mathbf{w}, \lambda)$, and $E_{\text{train}}(\mathbf{w}, \lambda)$:

	$E_{\text{train}}(\mathbf{w}, \lambda)$	$E_{\text{val}}(\mathbf{w}, \lambda)$
$\lambda_1 = 1$	0.51	0.55
$\lambda_2 = 0.1$	0.23	0.26
$\lambda_3 = 0.01$	0.05	0.42

   Which of the following statements are correct? **/1**

   [X] λ_1: underfitting, λ_2: best fit, λ_3: overfitting.

   [ ] λ_1: overfitting, λ_2: best fit, λ_3: overfitting.

   [ ] λ_1: overfitting, λ_2: best fit, λ_3: underfitting.

   [ ] λ_1: underfitting, λ_2: best fit, λ_3: underfitting.
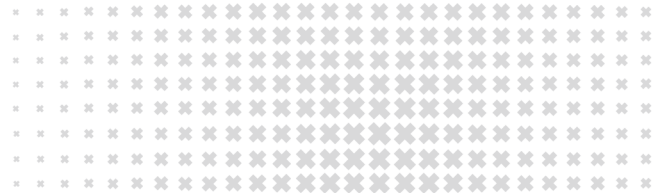
8. Consider two polynomial regression models, model 1 and model 2, of order M_1 and M_2 respecitvely, with $M_1 > M_2$. Which statements are correct? **/1**

- [X] If both models give a low training error, Model M_2 is more likely to lead to a low test error
- [ ] The training error is more likely to be lower for model 2 than for model 1
- [X] The training error is more likely to be lower for model 1 than for model 2.
- [X] Model 1 is more sensitive to overfitting than model 2.

9. We consider *maximum likelihood* (ML) and *maximum a posteriori* (MAP) optimization. Let $\mathcal{D}$ be the data and $\mathbf{w}$ the model parameters. Choose the correct statements: **/1**

- [X] For ML optimization we choose the model parameters as $\arg\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w})$.
- [ ] For MAP optimization we choose the model parameters as $\arg\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w})$
- [X] For MAP optimization we choose the model parameters as $\arg\max_{\mathbf{w}} p(\mathbf{w}|\mathcal{D})$
- [ ] MAP optimization is harder than ML optimization because we choose the model parameters as $\arg\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w})p(\mathbf{w})/p(\mathcal{D})$, and we need to know $p(\mathcal{D}) = \int \mathrm{d}\mathbf{w}\,p(\mathcal{D}, \mathbf{w})$ in order to find the position of the maximum.

10. Let $\mathcal{D}$ be the training dataset with an i.i.d. assumption on the data distribution, and $\mathbf{w}$ the model parameters. Which of the following statements about Bayesian linear regression are correct? **/1**

- [ ] The standard deviation of the predictive distribution for the target t' of a new datapoint $\mathbf{x}'$ is independent of $\mathbf{x}'$.
- [X] The predictive distribution for the target t' of a new datapoint $\mathbf{x}'$ takes into account the inherent noise of the true distribution of t', as well as the uncertainty in the values of $\mathbf{w}$.
- [X] The standard deviation of the predictive distribution decreases when $\mathbf{x}'$ becomes closer to one of inputs $\mathbf{x}_n \in D$.
- [ ] When the number of datapoints in the training dataset D becomes infinite, the standard deviation in the predictive distribution will go to zero.

11. You are given a dataset $\mathcal{D} = \{x_n\}_{n=1}^N$. The data is normally distributed $\mathcal{N}(x_n|\mu, \sigma^2)$ and we assume a Gaussian prior over μ : $\mathcal{N}(\mu|0, \sigma_0^2)$. Furthermore, the variance σ^2 is assumed to be known. Indicate which of the following statements are correct: **/1**

- [ ] When N (the number of datapoints in D) becomes larger, the prior distribution will become narrower.
- [ ] When N (the number of datapoints in D) becomes larger, the prior distribution will become wider.
- [X] When σ_0^2 is small, the maximum a posteriori estimate of μ will be strongly influenced by the prior.
- [X] When σ_0^2 is large, the maximum a posteriori estimate of μ will be only weakly influenced by the prior.
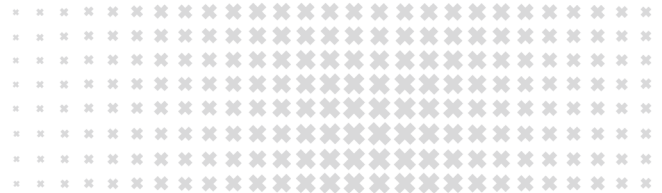
12. Given a data set $\mathcal{D} = \{x_n\}_{n=1}^N$, with each data point normally distributed by $\mathcal{N}(x_n|\mu, \beta^{-1})$. We assume a Gaussian prior over μ : $\mathcal{N}(\mu|0, \alpha^{-1})$. Furthermore, the precision β is assumed to be known. How does the posterior change as (i) $\alpha \to 0$, (ii) $\alpha \to \infty$ (iii) $N \to \infty$ **/1**

   ☐ (i) wider (ii) narrower (iii) same.

   ☒ (i) wider (ii) narrower (iii) narrower.

   ☐ (i) narrower (ii) wider (iii) narrower.

   ☐ (i) narrower (ii) wider (iii) same.

13. Let $\mathcal{D}$ be the training dataset with an i.i.d. assumption on the data distribution, and $\mathbf{w}$ the model parameters. Which of the following describes a Bayesian predictive distribution for the target variable t' of a new datapoint $\mathbf{x}'$? **/1**

   ☐ $p(t'|D, \mathbf{x}') = \int \int p(t|\mathbf{x}', \mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}\mathrm{d}\mathbf{x}'$.

   ☒ $p(t'|D, \mathbf{x}') = \int p(t', \mathbf{w}|\mathcal{D}, \mathbf{x}')\mathrm{d}\mathbf{w}$.

   ☒ $p(t'|D, \mathbf{x}') = \int p(t'|\mathbf{x}', \mathbf{w})p(\mathbf{w}|D)\mathrm{d}\mathbf{w}$.

   ☐ $p(t'|D, \mathbf{x}') = \int p(t|\mathbf{x}', \mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w}$.

14. In classification there are three approaches: discriminant functions, probabilistic generative models and probabilistic discriminative models. Indicate which statements are correct. **/1**

   ☒ Logistic regression is a probabilistic discriminative model.

   ☐ For datapoints $\mathbf{x} \in \mathbb{R}^D$, and no use of basis functions, logistic regression uses more parameters than Linear Discriminant Analysis.

   ☐ In discriminant function models, the probability of class $\mathcal{C}_k$ given input $\mathbf{x}$, given by $p(\mathcal{C}_k|\mathbf{x})$, is modeled directly.

   ☒ In Linear Discriminant Analysis, the class conditional densities $p(\mathbf{x}|\mathcal{C}_k)$ are modeled with Gaussian distributions.

   ☐ With probabilistic discriminative models new data points can be generated.

15. Consider stochastic gradient descent (SGD) for Logistic Regression with error function $E(\mathbf{w})$. Which of the following statements are correct? **/1**

   ☐ In order to avoid getting stuck in a local minimum of $E(\mathbf{w})$ (which is not equal to the global minimum), we should make sure the learning rate η is large enough.

   ☐ The stochastic gradient descent algorithm will always converge for logistic regression, no matter how you choose the learning rate η.

   ☐ If after a few iterations $E(\mathbf{w})$ increases instead of decreases then most likely the learning rate η is too small.

   ☒ The error function $E(\mathbf{w})$ for logistic regression is convex, so it has only one (global) minimum.

   ☒ Stochastic gradient descent is useful for Logistic Regression problems, because no closed-form (analytic) solution for the global minimum exists.

16. Consider SGD and the iterative reweighted least squares (IRLS) algorithm for Logistic Regression with error function $E(\mathbf{w})$. Which of the following statements are correct? **/1**

    ☐ SGD converges faster to the minimum of $E(\mathbf{w})$ than IRLS because SGD makes updates in the direction of steepest descent.

    ☒ IRLS is based on making a local quadratic approximation to $E(\mathbf{w})$ around the previous estimate of $\mathbf{w}$, and finding the next estimate of $\mathbf{w}$ as the minimizer of this approximation.

    ☐ The learning rate in IRLS can be chosen higher than is the case for SGD.

    ☐ For both IRLS and SGD we need to compute the gradient of $E(\mathbf{w})$ with respect to $\mathbf{w}$, as well as the Hessian matrix $\mathbf{H}$ with elements $H_{ij} = \frac{\partial E(\mathbf{w})}{\partial w_i \partial w_j}$.

General remarks

The solutions given below, with the corresponding distribution of points, serve as a guideline. If some intermediate steps are left implicit by the student, while still clearly following a derivation, points will not be deducted. The total number of possible points is 41, with 3 points as bonus points, meaning that the final grade is computed as $10 \times \frac{\#\text{points}}{38}$.

2   Probability Theory and Bayes' Rule

/8

Suppose one percent of all women over the age of 50 have breast cancer. Of all women over 50 who have breast cancer, 90 percent tests positive for cancer with a mammogram test. 6 percent of all women over 50 will have false positive mammograms.

a) Define all random variables and the values they can take. **/2**

b) What is the probability that a woman over the age of 50 has cancer if she had a positive mammogram result? You can leave your numerical answer in the form of a fraction if you do not have a calculator with you. **/6**

Solutions

a) $C \in \{c, h\}$ cancer status (cancer, no cancer/healthy) (1pt), $M \in \{+, -\}$ mammogram test outcome (positive for cancer, negative for cancer) (1pt).

b) In this exercise we only consider women over the age of 50. The probability of having cancer for those women is equal to $p(c) = 0.01$, and thus the probability of not having cancer is $p(h) = 0.99$ (1pt). The conditional probability of testing positive for cancer through a mammogram test, given that the patient actually has cancer is given by $p(+|c) = 0.9$ (1pt). The probability of testing positive for cancer while not having cancer is given by $p(+|h) = 0.06$(1pt). The probability to compute is $p(c|+)$. Using Bayes rule this results in (1pt)

$$p(c|+) = \frac{p(+|c)p(c)}{p(+)},$$

with $p(+)$ the probability of having a positive mammogram test: $p(+) = p(+, c) + p(+, h) = p(+|c)p(c) + p(+|h)p(h)$ (1pt). Inserting the numbers given above leads to (1pt)

$$p(c|+) = \frac{p(+|c)p(c)}{p(+|c)p(c) + p(+|h)p(h)} = \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.06 \times 0.99} \approx 0.13.$$

3   Decision boundaries for two-class Logistic Regression

/3

We fit a Logistic Regression model with two classes $t \in \{0, 1\}$ for two dimensional data points $\mathbf{x} = (x_1, x_2)$, with conditional class probabilities $p(t = 1|\mathbf{x}, \mathbf{w}) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$. After our learning algorithm has converged, we obtain an estimate for the model parameters $\mathbf{w}^* = (w_0^*, w_1^*, w_2^*)^T = (0, 1, -1)^T$. Which of the decision boundaries in the (x_1, x_2)-plane shown below corresponds to the trained model? Explain your answer.

(a)   (b)   (c)
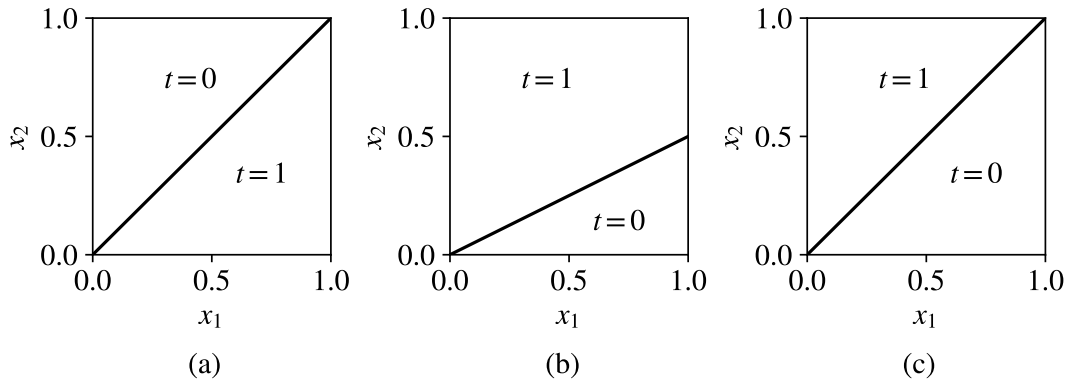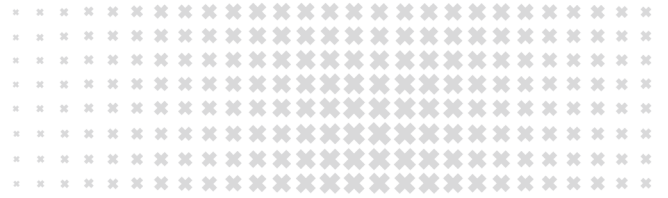
Solutions

Picture (a) represents the correct decision boundary. The decision boundary can be computed by solving $\sigma(w_0^* + w_1^* x_1 + w_2^* x_2) = 0.5$, or equivalently $w_0^* + w_1^* x_1 + w_2^* x_2 = 0$ (1 pt). Using $(w_0^*, w_1^*, w_2^*)^T = (0, 1, -1)^T$, we get $x_1 - x_2 = 0$ and the decision boundary is thus given by the line $x_1 = x_2$ (1 pt) , thus eliminating the possibility of picture (b) being correct. When $w_0^* + w_1^* x_1^* + w_2^* x_2 > 0$ the class $t = 1$ will be predicted, in this case given by $x_1 > x_2$, as shown in the picture (a) (1 pt). Note that simply plugging in points for (x_1, x_2), such as $(0, 1)$, and showing that the right class label comes out for picture (a), is also a valid way of solving this problem.

4 Regularized Logistic Regression for K classes

/6

Consider logistic regression for K classes with N training vectors $\{\mathbf{x}_n\}_{n=1}^N$, each of which is mapped to a different feature vector $\boldsymbol{\phi}(\mathbf{x}_n) = \boldsymbol{\phi}_n = (\phi_0(\mathbf{x}_n), \phi_1(\mathbf{x}_n), ..., \phi_{M-1}(\mathbf{x}_n))^T$ using basis functions $\phi_j(\mathbf{x})$ with $j = 0, ..., M-1$, and $\phi_0(\mathbf{x}) = 1$. Each vector $\mathbf{x}_n$ has a corresponding target vector $\mathbf{t}_n$ of size K: $\mathbf{t}_n = (t_{n1}, t_{n2}, ..., t_{nK})^T$, where $t_{nk} = 1$ if $\mathbf{x}_n \in \mathcal{C}_k$, and $t_{nj} = 0$ for all $j \neq k$. The input data can be collected in a matrix $\mathbf{X}$ with row n given by $\mathbf{x}_n^T$, and the targets can be collected in a target matrix $\mathbf{T}$, with row n equal to $\mathbf{t}_n^T$. The feature vectors can also be collected in a matrix $\boldsymbol{\Phi}$ such that the nth row of $\boldsymbol{\Phi}$ contains $\boldsymbol{\phi}_n^T$:

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

Assuming i.i.d. data, the posterior class probabilities are modeled by

$$p(\mathcal{C}_k | \boldsymbol{\phi}(\mathbf{x}), \mathbf{w}_1, ..., \mathbf{w}_K) = y_k(\boldsymbol{\phi}) = \frac{\exp(a_k(\boldsymbol{\phi}))}{\sum_{j=1}^K \exp(a_j(\boldsymbol{\phi}))},$$

where $a_k(\boldsymbol{\phi}) = \mathbf{w}_k^T \boldsymbol{\phi}$ with $\boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{x})$, and $\mathbf{w}_k = (w_{k0}, ..., w_{kM-1})^T$. Assume a Gaussian prior on the parameter vectors $\mathbf{w}_1, ..., \mathbf{w}_K$:

$$p(\mathbf{w}_1, ..., \mathbf{w}_K | \alpha) = \prod_{k=1}^K \mathcal{N}(\mathbf{w}_k | \mathbf{0}, \alpha^{-1}\mathbb{1}),$$

with $\mathbb{1}$ the identity matrix. Show that obtaining a Maximum A Posteriori (MAP) estimate for $\mathbf{w}_1, ..., \mathbf{w}_K$ is equivalent to performing regularized logistic regression for K classes, where we minimize the function

$$-\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_k(\boldsymbol{\phi}_n) + \frac{\alpha}{2}\sum_{k=1}^{K}\sum_{m=0}^{M-1} |w_{km}|^2$$

with respect to $\mathbf{w}_1, ..., \mathbf{w}_K$.

Solutions

The posterior distribution over the parameters is given by

$$p(\mathbf{w}_1, ..., \mathbf{w}_K | \boldsymbol{\Phi}, \mathbf{T}, \alpha) = \frac{p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{w}_1, ..., \mathbf{w}_K)p(\mathbf{w}_1, ..., \mathbf{w}_K|\alpha)}{p(\mathbf{T}|\boldsymbol{\Phi}, \alpha)}. \text{ (1 pt)}$$

The MAP estimate for $\mathbf{w}_1, ..., \mathbf{w}_K$ is obtained by maximizing the posterior distribution with respect to the parameters $\mathbf{w}_1, ..., \mathbf{w}_K$:

$$\mathbf{w}_1^{\text{MAP}}, ..., \mathbf{w}_K^{\text{MAP}} = \underset{\mathbf{w}_1,...,\mathbf{w}_K}{\arg\max}\, p(\mathbf{w}_1, ..., \mathbf{w}_K | \boldsymbol{\Phi}, \mathbf{T}, \alpha) = \underset{\mathbf{w}_1,...,\mathbf{w}_K}{\arg\min}\, -\ln p(\mathbf{w}_1, ..., \mathbf{w}_K | \boldsymbol{\Phi}, \mathbf{T}, \alpha)$$

$$= \underset{\mathbf{w}_1,...,\mathbf{w}_K}{\arg\min}\, -\ln p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{w}_1, ..., \mathbf{w}_K) - \ln p(\mathbf{w}_1, ..., \mathbf{w}_K|\alpha), \text{ (2 pt)}$$

where we have used the fact that $\frac{\partial}{\partial \mathbf{w}_j} - \ln p(\mathbf{T}|\boldsymbol{\Phi}, \alpha) = 0$ for $j = 1, ..., K$. The log-likelihood is given by

$$\ln p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{w}_1, ..., \mathbf{w}_K) = \ln \prod_{n=1}^{N}\prod_{k=1}^{K} y_k(\boldsymbol{\phi}_n)^{t_{nk}} = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_k(\boldsymbol{\phi}_n). \text{ (1 pt)}$$

The log of the prior is given by

$$\ln p(\mathbf{w}_1, ..., \mathbf{w}_K|\alpha) = \ln \prod_{k=1}^{K} \mathcal{N}(\mathbf{w}_k|\mathbf{0}, \alpha^{-1}\mathbb{1}) = -\frac{KM}{2}\ln 2\pi + \frac{KM}{2}\ln\alpha - \frac{\alpha}{2}\sum_{k=1}^{K} \mathbf{w}_k^T\mathbf{w}_k. \text{ (1 pt)}$$

Noting that the first two terms are independent of $\mathbf{w}_1, ..., \mathbf{w}_K$, and using $\mathbf{w}_k^T\mathbf{w}_k = \sum_{m=0}^{M-1}|w_{km}|^2$, we obtain

$$\mathbf{w}_1^{\text{MAP}}, ..., \mathbf{w}_K^{\text{MAP}} = \underset{\mathbf{w}_1,...,\mathbf{w}_K}{\arg\min}\, -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_k(\boldsymbol{\phi}_n) + \frac{\alpha}{2}\sum_{k=1}^{K}\sum_{m=0}^{M-1} |w_{km}|^2. \text{ (1 pt)}$$

5 Maximum Likelihood for linear regression with multiple outputs

/8

Consider linear regression with N training vectors $\{\mathbf{x}_n\}_{n=1}^{N}$, each of which is mapped to a different feature vector $\boldsymbol{\phi}_n = (\phi_0(\mathbf{x}_n), \phi_1(\mathbf{x}_n), ..., \phi_{M-1}(\mathbf{x}_n))^T$ using basis functions $\phi_j(\mathbf{x})$ with $j = 0, ..., M-1$ where we define $\phi_0(\mathbf{x}) = 1$. In the training set, the data comes in input-output pairs: $\{\mathbf{x}_n, \mathbf{t}_n\}$, where the targets $\mathbf{t}_n$ are now K-dimensional vectors: $\mathbf{t}_n = (t_{n1}, t_{n2}, ..., t_{nK})^T$. We also have the following information:

- The regression prediction is given by a K-dimensional vector: $\mathbf{y}(\mathbf{x}_n, \mathbf{W}) = \mathbf{W}^T \boldsymbol{\phi}_n$, where $\mathbf{W}$ is a matrix of parameters with M rows and K colums.

- The likelihood function is a multivariate Gaussian: $p(\mathbf{t}_n|\boldsymbol{\phi}_n, \mathbf{W}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\mathbf{t}_n|\mathbf{W}^T\boldsymbol{\phi}_n, \boldsymbol{\Sigma}\right)$

- The data are independently and identically distributed (i.i.d).

The input data can be collected in a matrix $\mathbf{X}$ with row n given by $\mathbf{x}_n^T$, and the targets can be collected in a target matrix $\mathbf{T}$, with row n equal to $\mathbf{t}_n^T$. The feature vectors can also be collected in a matrix $\boldsymbol{\Phi}$ such that the nth row of $\boldsymbol{\Phi}$ contains $\boldsymbol{\phi}_n^T$:

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

In case you need to invert a matrix in any of the following questions make the explicit assumption that the matrix is invertible.

a) Compute the likelihood $p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{W}, \boldsymbol{\Sigma})$ and the log-likelihood $\ln p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{W}, \boldsymbol{\Sigma})$. /2

b) Compute $\frac{\partial}{\partial W_{ij}} \ln p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{W}, \boldsymbol{\Sigma})$. It is easiest to use the following chain rule

$$\frac{\partial}{\partial W_{ij}} \ln p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{W}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \frac{\partial \ln p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{W}, \boldsymbol{\Sigma})}{\partial y_k(\mathbf{x}_n, \mathbf{W})} \frac{\partial y_k(\mathbf{x}_n, \mathbf{W})}{\partial W_{ij}}.$$

Use the following form of the log-likelihood:

$$\ln p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{W}, \boldsymbol{\Sigma}) = -\frac{NK}{2}\ln 2\pi - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W}))^T\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W})).$$

/3

c) Compute the maximum likelihood solution $\mathbf{W}_{\mathrm{ML}}$. Use the following result:

$$\frac{\partial}{\partial W_{ij}} \ln p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{W}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W}))^T\boldsymbol{\Sigma}^{-1}_{:,j}\phi_i(\mathbf{x}_n), \tag{1}$$

where $\boldsymbol{\Sigma}^{-1}_{:,j}$ denotes the j-th column of the matrix $\boldsymbol{\Sigma}^{-1}$. Hint: Using Eq. (1) construct an equation for $\mathbf{W}$, which you need to solve to obtain the maximum likelihood solution $\mathbf{W}_{\mathrm{ML}}$.

/3

Solutions

a) The likelihood is given by (1 pt)

$$p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{W}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \frac{1}{(2\pi)^{K/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W}))^T\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W}))\right].$$

The log likelihood is then given by (1 pt)

$$\ln p(\mathbf{T}|\boldsymbol{\Phi}, \mathbf{W}, \boldsymbol{\Sigma}) = -\frac{NK}{2}\ln 2\pi - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W}))^T\boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W})).$$

b)

$$\frac{\partial}{\partial W_{ij}} \ln p(\mathbf{T}|\mathbf{\Phi}, \mathbf{W}, \mathbf{\Sigma}) = \sum_{k=1}^{K} \frac{\partial \ln p(\mathbf{T}|\mathbf{\Phi}, \mathbf{W}, \mathbf{\Sigma})}{\partial y_k(\mathbf{x}_n, \mathbf{W})} \frac{\partial y_k(\mathbf{x}_n, \mathbf{W})}{\partial W_{ij}}$$

$$= -\frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} -2(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W}))^T \mathbf{\Sigma}_{:,k}^{-1} \frac{\partial y_k(\mathbf{x}_n, \mathbf{W})}{\partial W_{ij}}. \text{ (1 pt)}$$

Use

$$\frac{\partial y_k(\mathbf{x}_n, \mathbf{W})}{\partial W_{ij}} = \frac{\partial \left[\mathbf{W}^T \phi(\mathbf{x}_n)\right]_k}{\partial W_{ij}} = \frac{\partial}{\partial W_{ij}} \sum_{l=0}^{M-1} W_{lk} \phi_l(\mathbf{x}_n) = I[k=j] \phi_i(\mathbf{x}_n), \text{ (1 pt)}$$

with $I[k=j]$ the indicator function, which is equal to 1 if $k=j$, and 0 otherwise. This leads to the result

$$\frac{\partial}{\partial W_{ij}} \ln p(\mathbf{T}|\mathbf{\Phi}, \mathbf{W}, \mathbf{\Sigma}) = -\frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} -2(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W}))^T \mathbf{\Sigma}_{:,k}^{-1} I[k=j] \phi_i(\mathbf{x}_n)$$

$$= \sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W}))^T \mathbf{\Sigma}_{:,j}^{-1} \phi_i(\mathbf{x}_n). \text{ (1 pt)}$$

c) In order to obtain an equation that allows us to solve for $\mathbf{W}_{\mathrm{ML}}$, we need to set Eq. (1) equal to zero, and rewrite it into an equation for the matrix $\mathbf{W}$:

$$\frac{\partial}{\partial \mathbf{W}} \ln p(\mathbf{T}|\mathbf{\Phi}, \mathbf{W}, \mathbf{\Sigma}) = \sum_{n=1}^{N} \mathbf{\Sigma}^{-1} (\mathbf{t}_n - \mathbf{W}^T \phi(\mathbf{x}_n)) \phi(\mathbf{x}_n)^T = \mathbf{0}, \text{ (1 pt)}$$

where we have used $(\mathbf{a}\mathbf{b}^T)_{ij} = a_i b_j$ for two vectors $\mathbf{a}$ and $\mathbf{b}$, and $\mathbf{y}(\mathbf{x}_n, \mathbf{W}) = \mathbf{W}^T \phi(x_n)$. Furthermore, note that we have used the convention $[\partial f(\mathbf{X})/\partial \mathbf{X}]_{ij} = \partial f(\mathbf{X})/\partial X_{ji}$, together with the fact that $\mathbf{\Sigma}^{-1}$ is symmetric. Using the transposed convention is fine as well (next transpose step is then unnecessairy), as long as chain rules are applied correctly (in this case we use a component wise chain rule so it does not matter). Taking the transpose then yields

$$\sum_{n=1}^{N} \phi(\mathbf{x}_n)(\mathbf{t}_n^T - \phi(\mathbf{x}_n)^T \mathbf{W}) \mathbf{\Sigma}^{-1} = \mathbf{\Phi}^T (\mathbf{T} - \mathbf{\Phi} \mathbf{W}) \mathbf{\Sigma}^{-1} = \mathbf{0}.$$

Note that the last form with $\mathbf{\Phi}$ and $\mathbf{T}$, with the sum over n written as a matrix multiplication, is not necessary in order to obtain a correct answer below. Multiplying from the right with $\mathbf{\Sigma}$ we obtain (2 pt):

$$\mathbf{W}_{\mathrm{ML}} = \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{T}$$

or

$$\mathbf{W}_{\mathrm{ML}} = \left(\sum_{n=1}^{N} \phi_n \phi_n^T\right)^{-1} \sum_{n=1}^{N} \phi_n \mathbf{t}_n^T$$