Exam Preparation and HW7

Introduction to Machine Learning 2020

Julian Mäder

Schedule

- Exam 2019, Question 1
- Exam 2019, Question 2
- HW7 Questions 13, 14 and 15

Exam 2019, Question 1

This questions is about weighted linear regression. You are given a dataset consisting of n labeled training points $\mathcal{D} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.

In addition, you are given a set of non-negative weights $\{\lambda_1, \ldots, \lambda_n\}$, where $\sum_{i=1}^{n} \lambda_i = 1$. Each weight $\lambda_i \in \mathbb{R}_+$ reflects the importance of correctly estimating the label of a specific training point $(\mathbf{x}_i, y_i)$.

A common approach towards this task is to find a solution $\mathbf{w} \in \mathbb{R}^d$ which minimizes the *weighted empirical risk* $\hat{R}(\mathbf{w})$, which is defined as follows:

$$\hat{R}(\mathbf{w}) = \sum_{i=1}^{n} \lambda_i (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 .$$

Exam 2019, Question 1.1

(i) *Analytic Solution (MC)*
Let us denote by $\mathbf{X} \in \mathbb{R}^{n \times d}$ the matrix whose rows are $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\mathbf{y} \in \mathbb{R}^n$ a row vector whose entries are $\{y_1, \ldots, y_n\}$, and let $\mathbf{\Lambda} \in \mathbb{R}^{n \times n}$ be a diagonal matrix such $\mathbf{\Lambda}_{ii} = \lambda_i$

What is the closed form solution for the minimizer $\hat{\mathbf{w}} := \arg\min_{\mathbf{w} \in \mathbb{R}^d} \hat{R}(\mathbf{w})$?
Comment: You may assume that the matrices $\mathbf{X}^\top \mathbf{X}$ and $\mathbf{X}^\top \mathbf{\Lambda} \mathbf{X}$ are invertible.

- ☐ $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{\Lambda} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Lambda} \mathbf{y}$
- ☐ $\hat{\mathbf{w}} = \mathbf{\Lambda}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Lambda} \mathbf{y}$
- ☐ $\hat{\mathbf{w}} = \mathbf{\Lambda}^{1/2}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Lambda}^{1/2} \mathbf{y}$
- ☐ $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{\Lambda} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Lambda}^{1/2} \mathbf{y}$
- ☐ $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Lambda}^{1/2} \mathbf{y}$
- ☐ $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Lambda} \mathbf{y}$

Exam 2019, Question 1.1

$$\mathbf{X} = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & \dots & \dots \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nd} \end{pmatrix}, \quad \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ \dots \\ \dots \\ w_d \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \dots \\ \dots \\ y_n \end{pmatrix}$$

$$\hat{R}(\mathbf{w}) = \sum_{i=1}^{n} \lambda_i (\mathbf{w}^\top \mathbf{x}_i - y_i)^2 = \sum_{i=1}^{n} (\mathbf{w}^\top \mathbf{x}_i - y_i) \lambda_i (\mathbf{w}^\top \mathbf{x}_i - y_i)$$

$$= (\mathbf{X}\mathbf{w} - \mathbf{y})^\top \mathbf{\Lambda} (\mathbf{X}\mathbf{w} - \mathbf{y}) = (\mathbf{w}^\top \mathbf{X}^\top - \mathbf{y}^\top) \mathbf{\Lambda} (\mathbf{X}\mathbf{w} - \mathbf{y})$$

Exam 2019, Question 1.1

$$\hat{R}(\mathbf{w}) = (\mathbf{w}^\top \mathbf{X}^\top - \mathbf{y}^\top)\mathbf{\Lambda}(\mathbf{X}\mathbf{w} - \mathbf{y})$$

$$= \mathbf{w}^\top \mathbf{X}^\top \mathbf{\Lambda} \mathbf{X} \mathbf{w} - \mathbf{w}^\top \mathbf{X}^\top \mathbf{\Lambda} \mathbf{y} - \mathbf{y}^\top \mathbf{\Lambda} \mathbf{X} \mathbf{w} + \mathbf{y}^\top \mathbf{\Lambda} \mathbf{y}$$

$$\stackrel{*}{=} \mathbf{w}^\top \mathbf{X}^\top \mathbf{\Lambda} \mathbf{X} \mathbf{w} - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{\Lambda} \mathbf{y} + \mathbf{y}^\top \mathbf{\Lambda} \mathbf{y}$$

$* : \mathbf{y}^\top \mathbf{\Lambda} \mathbf{X} \mathbf{w}$ is a scalar

$$\Rightarrow \mathbf{y}^\top \mathbf{\Lambda} \mathbf{X} \mathbf{w} = (\mathbf{y}^\top \mathbf{\Lambda} \mathbf{X} \mathbf{w})^\top = \mathbf{w}^\top \mathbf{X}^\top \mathbf{\Lambda} \mathbf{y}$$

Exam 2019, Question 1.1

$$\hat{R}(\mathbf{w}) = \mathbf{w}^\top \mathbf{X}^\top \mathbf{\Lambda} \mathbf{X} \mathbf{w} - 2\mathbf{w}^\top \mathbf{X}^\top \mathbf{\Lambda} \mathbf{y} + \mathbf{y}^\top \mathbf{\Lambda} \mathbf{y}$$

$$\Rightarrow \nabla_w \hat{R}(\mathbf{w}) = 2\mathbf{X}^\top \mathbf{\Lambda} \mathbf{X} \mathbf{w} - 2\mathbf{X}^\top \mathbf{\Lambda} \mathbf{y} \stackrel{!}{=} 0$$

$$\Rightarrow \mathbf{X}^\top \mathbf{\Lambda} \mathbf{X} \hat{\mathbf{w}} = \mathbf{X}^\top \mathbf{\Lambda} \mathbf{y}$$

$$\Rightarrow \hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{\Lambda} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Lambda} \mathbf{y}$$

Exam 2019, Question 1.2

(ii) *Probabilistic Interpretation (MC)*
Consider the following probabilistic model. Assume that for all i,

$$y_i = \mathbf{w}^\top \mathbf{x}_i + \epsilon_i,$$

where $\mathbf{w} \in \mathbb{R}^d$ is a fixed (unknown) vector, and $\{\epsilon_1, \ldots, \epsilon_n\}$ are statistically independent Gaussian random variables such that

$$\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

where, $\sigma_i > 0$ is the standard deviation. The Maximum Likelihood Estimate (MLE) for this model is defined as follows,

$$\mathbf{w}_{\text{MLE}} := \arg\max_{\mathbf{w}} P(y_1, \ldots, y_n | x_1, \ldots x_n, \sigma_1, \ldots, \sigma_n, \mathbf{w}).$$

Exam 2019, Question 1.2

Recall that in class you have shown that if all σ_i's are the same then solving the above MLE problem is equivalent to minimizing the empirical risk $\arg\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top \mathbf{x}_i - y_i)^2$.

It can be shown that minimizing the weighted empirical risk appearing in the previous problem is equivalent to finding the MLE solution for an appropriate choice of $\sigma_1, \ldots, \sigma_n$. What should the relation be between σ_i and λ_i for this equivalence to hold?

- ☐ $\lambda_i \propto \sigma_i^{-2}$
- ☐ $\lambda_i \propto \sigma_i^{-1}$
- ☐ $\lambda_i \propto \sigma_i^{-1/2}$
- ☐ $\lambda_i \propto \sigma_i$
- ☐ $\lambda_i \propto \sigma_i^{1/2}$
- ☐ $\lambda_i \propto \sigma_i^{2}$
- ☐ $\lambda_i \propto \log(1 + \sqrt{\sigma_i})$

Exam 2019, Question 1.2

First we reformulate the maximum likelihood estimate:

$$\mathbf{w}_{MLE} = \underset{\mathbf{w}}{\operatorname{argmax}} P(y_1, ..., y_n | \mathbf{x}_1, ..., \mathbf{x}_n, \sigma_1, ..., \sigma_n, \mathbf{w})$$

$$\overset{i.i.d.}{=} \underset{\mathbf{w}}{\operatorname{argmax}} \prod_{i=1}^{n} P(y_i | \mathbf{x}_i, \sigma_i, \mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmax}}\, log \prod_{i=1}^{n} P(y_i | \mathbf{x}_i, \sigma_i, \mathbf{w})$$

$$= \underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{n} log P(y_i | \mathbf{x}_i, \sigma_i, \mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmin}} - \sum_{i=1}^{n} log P(y_i | \mathbf{x}_i, \sigma_i, \mathbf{w})$$

Exam 2019, Question 1.2

Next we look at our assumptions about the data:

We assume that $y_i = \mathbf{w}^\top \mathbf{x}_i + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$

$$\Rightarrow P(y_i|\mathbf{x}_i, \sigma_i, \mathbf{w}) = \mathcal{N}(y_i, \mathbf{w}^\top \mathbf{x}_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{(\mathbf{w}^\top \mathbf{x}_i - y_i)^2}{2\sigma_i^2})$$

And from the last slide we know $\mathbf{w}_{MLE} = \text{argmin}_{\mathbf{w}} - \sum_{i=1}^n log P(y_i|\mathbf{x}_i, \sigma_i, \mathbf{w})$

Exam 2019, Question 1.2

$$\mathbf{w}_{MLE} = \underset{\mathbf{w}}{\operatorname{argmin}} - \sum_{i=1}^{n} logP(y_i|\mathbf{x}_i, \sigma_i, \mathbf{w})$$

$$= \underset{\mathbf{w}}{\operatorname{argmin}} - \sum_{i=1}^{n} log(\frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{(\mathbf{w}^\top \mathbf{x}_i - y_i)^2}{2\sigma_i^2}))$$

$$= \underset{\mathbf{w}}{\operatorname{argmin}} - \sum_{i=1}^{n} (log(\frac{1}{\sqrt{2\pi}\sigma_i}) - \frac{(\mathbf{w}^\top \mathbf{x}_i - y_i)^2}{2\sigma_i^2})$$

$$= \underset{\mathbf{w}}{\operatorname{argmin}} - \sum_{i=1}^{n} \frac{1}{2\sigma_i^2} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2$$

Exam 2019, Question 1.2

So let's compare the maximum likelihood estimate to the weighted empirical risk:

$$\mathbf{w}_{MLE} = \operatorname{argmin}_{\mathbf{w}} - \sum_{i=1}^{n} \frac{1}{2\sigma_i^2} (\mathbf{w}^\top \mathbf{x}_i - y_i)^2$$

$$\hat{R}(\mathbf{w}) = \sum_{i=1}^{n} \lambda_i (\mathbf{w}^\top \mathbf{x}_i - y_i)^2$$

$$\Rightarrow \lambda_i \propto \sigma_i^{-2}$$

Exam 2019, Question 1.3

In order to improve generalization properties of our model, we introduce a regularization term to the training objective (same weights). This is especially beneficial when you have little data. The cost function becomes,

$$\hat{R}_\eta(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \eta C(\mathbf{w}).$$

Two common candidates seen in the course are L_1 (Lasso) and L_2 (Ridge) regularization. These correspond to $C_1(w) = ||\mathbf{w}||_1$, and $C_2(w) = ||\mathbf{w}||_2^2$ in the above formula (in place of C) respectively.

Exam 2019, Question 1.3

(iii) *Analytic solution for L_1 (MC)*

Please choose which of the following formulas corresponds to the closed form of the minimizer of the $\hat{R}_\eta(\mathbf{w})$ with $C(\mathbf{w}) = ||\mathbf{w}||_1$,

☐ $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X} + \eta \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$

☐ $\hat{\mathbf{w}} = (\eta \mathbf{I})^{1/2} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\eta \mathbf{I})^{1/2} \mathbf{y}$

☐ $\hat{\mathbf{w}} = (\mathbf{X}^\top (\mathbf{I} + \eta \mathbf{I}) \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$

☒ In general, there is no closed form. ✓

...because $||\mathbf{w}||_1$ is not differentiable!

Exam 2019, Question 1.4

$$\hat{R}_\eta(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \lambda_i(\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \eta C(\mathbf{w})$$

with $C(\mathbf{w}) = \|\mathbf{w}\|_1$ (Lasso) or $C(\mathbf{w}) = \|\mathbf{w}\|_2^2$ (Ridge)

(iv) *Regularization limits (T/F)*

Decide whether the following statements are true or false when $\eta \to \infty$:

True	False	
☒	☐	When $C(\mathbf{w}) = \|\|\mathbf{w}\|\|_1$, then the solution $\|\|\hat{\mathbf{w}}\|\|_2 \to 0$.
☐	☒	When $C(\mathbf{w}) = \|\|\mathbf{w}\|\|_1$, the regularization has no longer any effect on $\hat{w}$.
☐	☒	When $C(\mathbf{w}) = \|\|\mathbf{w}\|\|_1$ or $C(\mathbf{w}) = \|\|\mathbf{w}\|\|_2^2$ the solution $\|\|\hat{\mathbf{w}}\|\|_2 \to \infty$.
☐	☒	When $C(\mathbf{w}) = \|\|\mathbf{w}\|\|_2^2$ the regularization has no longer any effect on $\hat{w}$.

Exam 2019, Question 1.5

(v) *Different L_2 regularization (T/F)*

Suppose we use the regularizer $C(\mathbf{w}) = ||\mathbf{w}||_2^2$ and optimize $\hat{R}_{\eta_1}$ with a regularization constant η_1 to get the minimizer $\hat{\mathbf{w}}_1$, and $\hat{R}_{\eta_2}$ with a regularization constant η_2 to get the minimizer $\hat{\mathbf{w}}_2$.

We know that η_2 and η_1 are *arbitrary* and *positive*, and crucially,

$$\eta_2 > \eta_1.$$

Decide which of the following statements are true or false for all possible datasets $\{\mathbf{x}_i, y_i\}_{i=1}^n$:

True	False	
☒	☐	$\lvert\lvert\hat{\mathbf{w}}_2\rvert\rvert_2 \leq \lvert\lvert\hat{\mathbf{w}}_1\rvert\rvert_2$
☐	☒	The solution $\hat{\mathbf{w}}_2$ is sparser than $\hat{\mathbf{w}}_1$
☐	☒	Solutions are the same, i.e. $\hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_2$
☐	☒	There always exist η_1, η_2 s.t. $\eta_1 \neq \eta_2$ and $\hat{\mathbf{w}}_1 = \hat{\mathbf{w}}_2$.

Recap Kernels

Perceptron:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \max\{0, -y_i \mathbf{w}^T \mathbf{x}_i\}$$

Fundamental insight: Optimal hyperplane lies in the span of the data

$$\hat{\mathbf{w}} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

for some $\alpha_{1:n} \in \mathbb{R}^n$

$$\Rightarrow \; \ldots \; \Rightarrow \qquad \hat{\alpha} = \arg\min_{\alpha_{1:n}} \frac{1}{n} \sum_{i=1}^{n} \max\{0, -\sum_{j=1}^{n} \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j\}$$

$\mathbf{x}_i^T \mathbf{x}_j \Downarrow k(x_i, x_j)$

The „Kernel Trick“

- Express problem s.t. it only depends on inner products
- Replace inner products by kernels

$$\mathbf{x}_i^T \mathbf{x}_j \Rightarrow k(\mathbf{x}_i, \mathbf{x}_j)$$

Recap Kernels

Often $k(\mathbf{x}, \mathbf{x}')$ can be computed much more efficiently than $\phi(\mathbf{x})^\top \phi(\mathbf{x}')$.
Here is a simple example of a polynomial kernel of degree 2:

Feature transformation: $\mathbf{x} = (x_1, x_2) \mapsto \phi(\mathbf{x}) := (x_1^2, x_2^2, \sqrt{2} x_1 x_2)$

Not kernelized: $\mathbf{x}^\top \mathbf{x}' \mapsto \phi(\mathbf{x})^\top \phi(\mathbf{x}') = x_1^2 x_1'^2 + x_2^2 x_2'^2 + 2 x_1 x_2 x_1' x_2'$

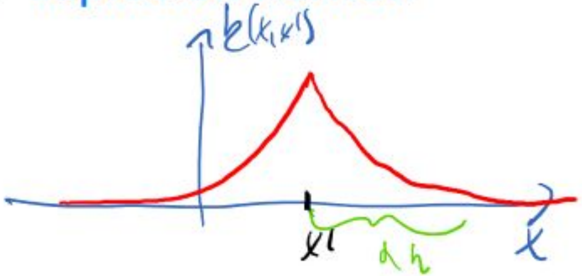
Kernelized: $\mathbf{x}^\top \mathbf{x}' \mapsto k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}')^2 = (x_1 x_1' + x_2 x_2')^2$

Examples of kernels on $\mathbb{R}^d$

- Linear kernel: $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$
- Polynomial kernel: $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}' + 1)^d$
- Gaussian (RBF, squared exp. kernel): $k(\mathbf{x}, \mathbf{x}') = \exp(-||\mathbf{x} - \mathbf{x}'||_2^2 / h^2)$

"Bandwidth"/ Length scale parameter

- Laplacian kernel: $k(\mathbf{x}, \mathbf{x}') = \exp(-||\mathbf{x} - \mathbf{x}'||_1 / h)$



Exam 2019, Question 2.1

(i) *Kernelization (T/F)*

Which of the following learning algorithms can be kernelized?

True	False		
☒	☐	Principal component analysis	→ Lecture Slides: Dimensionality Reduction //, slides 6 - 12
☒	☐	Logistic regression	→ Lecture Slides: Kernels //, slides 34 - 37
☒	☐	K-Means Clustering	→ Lecture Slides: Dimensionality Reduction //, slides 6 - 13
☒	☐	Nearest Neighbour Classification	→ See Kernel Nearest-Neighbor Algorithm, Yu et al. 2002

Exam 2019, Question 2.2

(ii) *Feature Maps (T/F)*
From the lectures, we know that every kernel admits a feature representation in an inner product space such that the kernel can be represented as inner product (for example; if the inner product is in the Euclidean space, $k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^\top \phi(\mathbf{y})$). Decide whether the following statements are true or false.

True	False	
☐	☒	The feature map ϕ induced by a kernel k is always one-to-one.
☒	☐	The identity map $\phi(x) = x$ defines the linear kernel.
☒	☐	The dimension of the Euclidean feature map ϕ induced by the cubic kernel $k(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^\top \mathbf{y})^3$ where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ grows at least at a polynomial rate in d.
☒	☐	The radial basis function kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|_2^2)$ has an infinite-dimensional feature map ϕ.

Exam 2019, Question 2.2

True	False	
☐	☒	The feature map ϕ induced by a kernel k is always one-to-one.

Consider the feature map $\mathbf{x} = (x_1, x_2) \mapsto \phi(\mathbf{x}) := (x_1^2, x_2^2, \sqrt{2}x_1x_2)$,

induced by the kernel $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}')^2$.

Therefore, the points $\mathbf{x}^1 = (1, 1)$ and $\mathbf{x}^2 = (-1, -1)$ are transformed to

$\phi(\mathbf{x}^1) = (1, 1, \sqrt{2}) = \phi(\mathbf{x}^2)$.

Exam 2019, Question 2.2

True	**False**	
☒	☐	The dimension of the Euclidean feature map ϕ induced by the cubic kernel $k(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^\top \mathbf{y})^3$ where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ grows at least at a polynomial rate in d.

The feature map induced by this kernel
$\mathbf{x} = (x_1, ..., x_d) \mapsto \phi(\mathbf{x}) := (1, x_1, ..., x_d, x_1^2, ..., x_d^2, x_1^3, ..., x_d^3, x_1 x_2, x_1 x_3, ...)$
contains all monomials up to degree 3 in d variables.

The number of monomials up to degree n in d variables is given by: $\binom{d+n}{n} = \frac{(d+n)!}{n!d!}$

$\Rightarrow$ In our case: $\binom{d+n}{n} = \frac{(d+3)!}{6d!} = \frac{(d+3)(d+2)(d+1)d!}{6d!} = \frac{(d+3)(d+2)(d+1)}{6}$

$\Rightarrow$ Growth rate: $\mathcal{O}(d^3)$

Kernel Definition

A **kernel** is a function $k : X \times X \to \mathbb{R}$ satisfying

1) Symmetry: For any $\mathbf{x}, \mathbf{x}' \in X$ it must hold that

$k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$

2) Positive semi-definiteness: For any *n*, any set $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subseteq X$, the kernel (Gram) matrix

$$\mathbf{K} = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \ldots & k(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \ldots & k(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix}$$

must be positive semi-definite

Kernel Definition

- Kernel function $k : X \times X \to \mathbb{R}$
- Take any finite subset of data $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subseteq X$
- Then the kernel (gram) matrix

$$\mathbf{K} = \begin{pmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \ldots & k(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \ldots & k(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix} = \begin{pmatrix} \phi(\mathbf{x}_1)^T \phi(\mathbf{x}_1) & \ldots & \phi(\mathbf{x}_1)^T \phi(\mathbf{x}_n) \\ \vdots & & \vdots \\ \phi(\mathbf{x}_n)^T \phi(\mathbf{x}_1) & \ldots & \phi(\mathbf{x}_n)^T \phi(\mathbf{x}_n) \end{pmatrix}$$

is positive semidefinite

Because $\mathbf{K} = \mathbf{\Phi}^\top \mathbf{\Phi}$ with $\mathbf{\Phi} = (\phi(\mathbf{x_1}), ..., \phi(\mathbf{x_n}))$

$$\Rightarrow \forall \mathbf{x} \in \mathbb{R}^n : \mathbf{x}^\top \mathbf{K} \mathbf{x} = \mathbf{x}^\top \mathbf{\Phi}^\top \mathbf{\Phi} \mathbf{x} = (\mathbf{\Phi} \mathbf{x})^\top (\mathbf{\Phi} \mathbf{x}) \geqslant 0$$

Kernel Rules

Suppose we have two kernels

$$k_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \qquad\qquad k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

defined on data space *X*

Then the following functions are valid kernels:

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$$

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}')\, k_2(\mathbf{x}, \mathbf{x}')$$

$$k(\mathbf{x}, \mathbf{x}') = c\, k_1(\mathbf{x}, \mathbf{x}') \text{ for } c > 0$$

$$k(\mathbf{x}, \mathbf{x}') = f(k_1(\mathbf{x}, \mathbf{x}'))$$

where *f* is a polynomial with positive coefficients or the exponential function

Exam 2019, Question 2.3

(iii) *Valid Kernels (T/F)*

Let $x, y \in \mathbb{R}$. Let $k_1(x, y)$ and $k_2(x, y)$ be any valid kernel functions on $\mathbb{R} \times \mathbb{R}$. Consider the definitions of the function $f(x, y)$ below. For which of these definitions is f always a valid kernel (True)?

Hint: $\cos(x + y) = \cos(x)\cos(y) - \sin(x)\sin(y)$

True	**False**	
☐	☒	$f(x, y) = ck_1(x, y)^2 k_2(x, y)$ for any $c \in \mathbb{R}$
☒	☐	$f(x, y) = \cos(x - y)$
☐	☒	$f(x, y) = \frac{1}{k_1(x,y)}$ assuming $k_1(x, y) > 0$ for all $x, y \in \mathbb{R}$
☒	☐	$f(x, y) = (k_1(x, y) + k_2(x, y))^2$

Exam 2019, Question 2.3

True	False	
☐	☒	$f(x, y) = ck_1(x, y)^2 k_2(x, y)$ for any $c \in \mathbb{R}$

Because c needs to be bigger than Zero!

True	False	
☒	☐	$f(x, y) = (k_1(x, y) + k_2(x, y))^2$

$$= k_1(x, y)k_1(x, y) + 2k_1(x, y)k_2(x, y) + k_2(x, y)k_2(x, y)$$

Exam 2019, Question 2.3

Hint: $\cos(x+y) = \cos(x)\cos(y) - \sin(x)\sin(y)$

True **False**

☒ ☐ $f(x,y) = \cos(x-y)$

$$= cos(x)cos(-y) - sin(x)sin(-y) = cos(x)cos(y) + sin(x)sin(y)$$

$\Rightarrow f(x,y)$ is symmetric

and $\phi(x) = (cos(x), sin(x))$ is the induced feature map.

$\Rightarrow f(x,y)$ is a valid kernel.

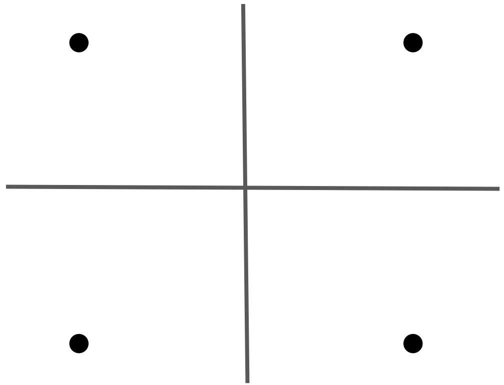
Exam 2019, Question 2.4

(iv) *Separable space (T/F)*

Consider a dataset consisting of the following four points in $\mathbb{R}^2$: $\mathbf{x}^1 = [-1,-1]^\top, \mathbf{x}^2 = [-1,1]^\top, \mathbf{x}^3 = [1,-1]^\top, \mathbf{x}^4 = [1,1]^\top$. Class labels for each point are unknown, but assume that each point $\mathbf{x}^i$ may belong to either of only two classes. You apply a feature transformation $\Phi(\cdot)$ to each point. For which of the feature transformations below is the resulting dataset $\{\Phi(\mathbf{x}^1), \Phi(\mathbf{x}^2), \Phi(\mathbf{x}^3), \Phi(\mathbf{x}^4)\}$ guaranteed to be linearly separable (with no point lying exactly on the decision boundary) for every possible class labelling (True)?

Hint: Note that subscript denotes the coordinate in this question, and superscript identifies the datapoint in the dataset.

True	False	
☐	☒	$\Phi(\mathbf{x}) = [\mathbf{x}_1, \mathbf{x}_2, 1]$
☐	☒	$\Phi(\mathbf{x}) = [\mathbf{x}_1^2, \mathbf{x}_2^2, \mathbf{x}_1, \mathbf{x}_2, 1]$
☐	☒	$\Phi(\mathbf{x}) = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_1^2 + \mathbf{x}_2^2, 1]$
☒	☐	$\Phi(\mathbf{x}) = [\mathbf{x}_1^2, \mathbf{x}_2^2, \mathbf{x}_1\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_2, 1]$



Exam 2019, Question 2.4

$\mathbf{x}^1 = [-1, -1]^\top, \mathbf{x}^2 = [-1, 1]^\top, \mathbf{x}^3 = [1, -1]^\top, \mathbf{x}^4 = [1, 1]^\top$

True	**False**	
☐	☒	$\Phi(\mathbf{x}) = [\mathbf{x}_1^2, \mathbf{x}_2^2, \mathbf{x}_1, \mathbf{x}_2, 1]$

For all 4 points it holds that: $\mathbf{\Phi}(\mathbf{x}^i) = [1, 1, \mathbf{x}_1, \mathbf{x}_2, 1]$

$\Rightarrow$ Still not linearly separable!

Exam 2019, Question 2.4

$\mathbf{x}^1 = [-1, -1]^\top, \mathbf{x}^2 = [-1, 1]^\top, \mathbf{x}^3 = [1, -1]^\top, \mathbf{x}^4 = [1, 1]^\top$

True	**False**	
☐	☒	$\Phi(\mathbf{x}) = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_1^2 + \mathbf{x}_2^2, 1]$

For all 4 points it holds that: $\boldsymbol{\Phi}(\mathbf{x}^i) = [\mathbf{x}_1, \mathbf{x}_2, 2, 1]$

$\Rightarrow$ Still not linearly separable!

Exam 2019, Question 2.4

$\mathbf{x}^1 = [-1,-1]^\top, \mathbf{x}^2 = [-1,1]^\top, \mathbf{x}^3 = [1,-1]^\top, \mathbf{x}^4 = [1,1]^\top$

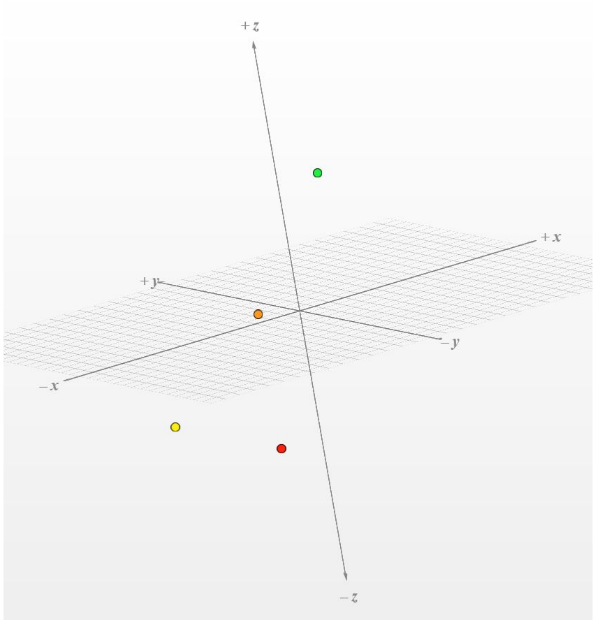
True	False	
☒	☐	$\Phi(\mathbf{x}) = [\mathbf{x}_1^2, \mathbf{x}_2^2, \mathbf{x}_1\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_2, 1]$

For all 4 points it holds that: $\boldsymbol{\Phi}(\mathbf{x}^i) = [1, 1, \mathbf{x}_1\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_2, 1]$

$\Rightarrow$ Look at: $\boldsymbol{\Phi}'(\mathbf{x}^i) = [\mathbf{x}_1\mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_2]$

$\Rightarrow \boldsymbol{\Phi}'(\mathbf{x}^1) = [-1,-1,-1], \boldsymbol{\Phi}'(\mathbf{x}^2) = [-1,-1,1],$
$\boldsymbol{\Phi}'(\mathbf{x}^3) = [-1,1,-1], \boldsymbol{\Phi}'(\mathbf{x}^4) = [1,1,1]$

$\Rightarrow$ Linearly separable!



Exam 2019, Question 2.5

(v) *Decision Boundaries (Matching Question)*
You have fitted the following four models to learn a classifier for a multi-class classification problem with three classes:

A. SVM with kernel $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$

B. SVM with kernel $k(\mathbf{x}, \mathbf{y}) = (\gamma \mathbf{x}^\top \mathbf{y} + 1)^3$

C. SVM with kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$

D. Nearest neighbour classifier (with five neighbours and uniform weighting)

All SVMs use a *one-vs-one* approach for the multi-class classification and are fitted using the same value for γ. The four figures below show the samples used for fitting all models and the decision boundaries generated by each classifier. Match each model above with its corresponding figures.

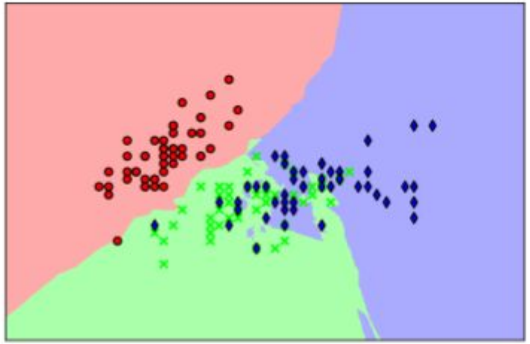
Exam 2019, Question 2.5

A. SVM with kernel $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$

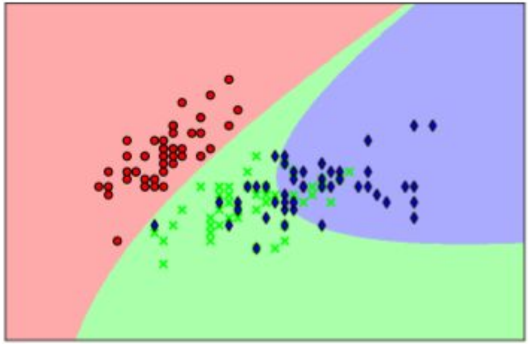
B. SVM with kernel $k(\mathbf{x}, \mathbf{y}) = (\gamma \mathbf{x}^\top \mathbf{y} + 1)^3$

C. SVM with kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$

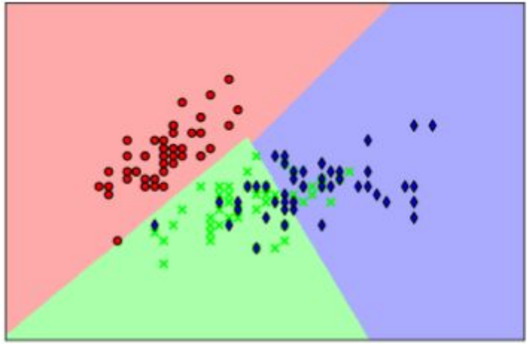
D. Nearest neighbour classifier (with five neighbours and uniform weighting)



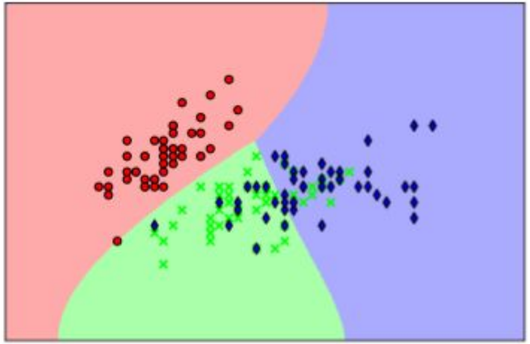
D. Nearest neighbour classifier



B. SVM with kernel $k(\mathbf{x}, \mathbf{y}) = (\gamma \mathbf{x}^\top \mathbf{y} + 1)^3$



A. SVM with kernel $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$



C. SVM with kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$

Exam 2019, Question 2.6

(vi) Consider the following function over real-valued scalars x and y:

$$k(x, y) = (1 + cxy)^2,$$

where c is a positive constant. The basis function of this kernel represent the kernel as $k(x, y) = \phi(x)^\top \phi(y)$, where $\phi(x) \in \mathbb{R}^3$. Given that $\phi(x) = [1, \star, cx^2]$, derive the expression that falls under the star.

$$\phi(x)^\top \phi(y) = [1, \star_x, cx^2]^\top [1, \star_y, cy^2] = 1 + \star_x \star_y + c^2 x^2 y^2$$

$$k(x, y) = (1 + cxy)^2 = 1 + 2cxy + c^2 x^2 y^2$$

$$\phi(x)^\top \phi(y) \stackrel{!}{=} k(x, y) \quad \Rightarrow \quad \star_x \star_y = 2cxy \quad \Rightarrow \quad \star_x = \sqrt{2c}x, \star_y = \sqrt{2c}y$$

$$\Rightarrow \phi(x) = [1, \sqrt{2c}x, cx^2]$$

Coffee Break

It's time for a coffee break, let's have a cup of coffee.

We'll Come Back After 15 Minutes

HW7 Question 13-15: Important Tipps

Expectation of a (discrete) Random Variable

Let X be a random variable with a finite number of finite outcomes $x_1, x_2, \ldots, x_k$ occurring with probabilities $p_1, p_2, \ldots, p_k$, respectively. The **expectation** of X is defined as

$$\mathrm{E}[X] = \sum_{i=1}^{k} x_i\, p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k.$$

HW7 Question 13-15: Important Tipps

Jensen's Inequality

If f is a convex function, we have

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Note that if X is constant we get an equality. Suppose we have $f(x) = x^2$, which is a convex function. Then, using Jensen's Inequality, we have $(\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$, which you may recall from the definition of $\mathrm{Var}(X)$. Moreover, if f is a concave function (e.g. $f(x) = \log x$), we reverse the inequality sign.

HW7 Question 13

In this question you will show that EM can be seen as an iterative algorithm which maximizes a lower bound on the log-likelihood. We will treat any general model $P(X, Z)$ with observed variables X and latent variable Z. For the sake of simplicity, we will assume that Z is discrete and takes values in $1, 2, ..., m$. If we observe X, the goal is to maximize the log-likelihood

$$l(\theta) = logP(\mathbf{x}; \theta) = log\Sigma_{z=1}^{m} P(\mathbf{x}, z; \theta)$$

with respect to the parameter vector θ. $Q(Z)$ denotes any distribution over the latent variables.

13. For $Q(z) > 0$ when $P(\mathbf{x}, z) > 0$, find a lower bound for the likelihood, $l(\theta)$. Hint: Consider using the Jensen's inequality.

    (a) $\mathbb{E}_Q[logP(X, Z)] - \Sigma_{z=1}^{m} Q(z) logQ(z)$
    (b) $\mathbb{E}_Q[logP(X, Z)] + \Sigma_{z=1}^{m} Q(z) logQ(z)$
    (c) $\mathbb{E}_Q[logP(X, Z)]$
    (d) $\mathbb{E}_Q[logP(X, Z)] + \Sigma_{z=1}^{m} Q(\mathbf{x}) logQ(\mathbf{x})$

HW7 Question 13

$$l(\theta) = log\sum_z P(x,z;\theta) = log\sum_z \frac{P(x,z;\theta)}{Q(z)}Q(z) \overset{*}{=} log\mathbb{E}_{Z\sim Q}[\frac{P(x,z;\theta)}{Q(z)}]$$

$$\overset{**}{\geqslant} \mathbb{E}_{Z\sim Q}[log\frac{P(x,z;\theta)}{Q(z)}] = \mathbb{E}_{Z\sim Q}[logP(x,z;\theta) - logQ(z)]$$

$$= \mathbb{E}_{Z\sim Q}[logP(x,z;\theta)] - \mathbb{E}_{Z\sim Q}[logQ(z)]$$

$$\overset{*}{=} \mathbb{E}_{Z\sim Q}[logP(x,z;\theta)] - \sum_z Q(z)logQ(z) \Rightarrow \text{(a) is the correct answer!}$$

* Expectation ** Jensen's Inequality

HW7 Question 14

For a fixed θ, pick the distribution $Q^*(Z)$ which maximizes the lower bound derived in the previous question. Show by yourself that bound is exact for this specific distribution. Hint: Do not forget to add Lagrange multipliers to make sure that Q^* is a valid distribution.

(a) $P(Z|\mathbf{x};\theta)$

(b) $P(Z;\theta)$

(c) $P(\mathbf{X}|z;\theta)$

(d) $P(\mathbf{X},Z;\theta)$

HW7 Question 14

We start with: $log\mathbb{E}_{Z\sim Q}[\frac{P(x,z;\theta)}{Q(z)}] \overset{**}{\geqslant} \mathbb{E}_{Z\sim Q}[log\frac{P(x,z;\theta)}{Q(z)}]$

We know, from Jensens Inequality, that the equality holds if $\frac{P(x,z;\theta)}{Q(z)}$ is constant.

$\Rightarrow Q^*(z) = cP(x,z;\theta)$

For some constant c that does not depend on z.

* Expectation ** Jensen's Inequality

HW7 Question 14

Additionally, we know that $Q^*(z)$ has to be a valid distribution: $\sum_z Q^*(z) = 1$

$$\Rightarrow Q^*(z) = cP(x,z;\theta) = \frac{cP(x,z;\theta)}{\sum_z Q^*(z)} = \frac{cP(x,z;\theta)}{\sum_z cP(x,z;\theta)} = \frac{P(x,z;\theta)}{\sum_z P(x,z;\theta)} = \frac{P(x,z;\theta)}{P(x;\theta)} = P(Z|x;\theta)$$

$\Rightarrow$ (a) is the correct answer!

HW7 Question 15

Mark the following statements True or False.

(a) Optimizing the lower bound on likelihood with respect to $Q(.)$ is exactly the E-step. ✓

(b) Optimizing the lower bound on likelihood with respect to $Q(.)$ is exactly the M-step.

(c) Optimizing the lower bound on likelihood with respect to θ for fixed $Q(.)$ is exactly the E-step.

(d) Optimizing the lower bound on likelihood with respect to θ for fixed $Q(.)$ is exactly the M-step. ✓

(e) The lower bound on likelihood monotonically increases after each step of optimisation. ✓

(f) The lower bound on likelihood monotonically decreases after each step of optimisation.

There is a more detailed explanation in the CS229 lecture notes (Part IX, The EM Algorithm) by Andrew Ng:
(https://course.ccs.neu.edu/cs6220f16/sec3/assets/pdf/cs229-notes8.pdf)