# Machine Learning 2020-21
# Final Exam

15 December 2020

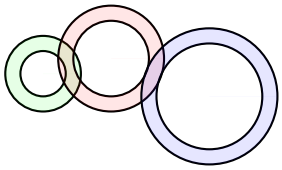
Name: .....

**Question 1:** 2 points In linear regression with L2 regularization, the augmented training loss is

$$L_{\text{aug}}(w) = \frac{1}{m}(w^\top \mathrm{X}^\top \mathrm{X} w - 2w^\top \mathrm{X}^\top y + y^\top y + \lambda w^\top w),$$

and the optimal weight vector is $w_{\text{reg}} = (\mathrm{X}^\top \mathrm{X} + \lambda I)^{-1} \mathrm{X}^\top y$. What is the augmented training loss of $w_{\text{reg}}$? Show the entire derivation and simplify the expression as much as possible.

**Question 2:** 2 points A company develops an application that estimates how long it will take users to drive to work. To do so, the company collects data of users driving to work, recording the initial and final location of the drive, and the time of day at the beginning and at the end. Which type of machine learning problem is this, and how are the concrete components of this learning problem defined?



**Question 3:** 1 point Consider an input space in two dimensions in which the true concept classes are defined as (possibly overlapping) circular bands, as shown in the figure. Now imagine that you sample many data points from this input space, such that each data point falls within one of the circular bands. However, you do not have access to labels. If you wanted to perform unsupervised learning for this problem, which concrete algorithm would you choose, and why?

**Question 4:** 1 point How is the bias-variance tradeoff defined? Give an example of a learning problem with low bias, and a learning problem with low variance, and use the examples to discuss the tradeoff.

**Question 5:** 1 point What is the meaning of the "kernel trick"? Describe how the kernel trick is applied in machine learning, and why it is important.

Name: .....

**Question 6:** 2 point Given some dataset $\mathcal{D}$ and some model hypothesis parameterized by parameters $\theta$, explain three main differences between estimating $\theta$ using Bayesian estimation and maximum likelihood. Illustrate the Bayesian approach with an example.

**Question 7:** 2 point We want to learn a generative model of text from a large corpus comprising senteces of different length in natural language. As a first approximation, we assume that the probability of a letter $X_k$ depends essentially on the two letters that appeared previously in the text, $X_{k-1}$ and $X_{k-2}$. Describe a Bayesian network representing such a model and then transform it into a factor graph. Formalize two conditional independencies that are derived from such a model.

Name: .....

**Question 8:** [1 point] In discounted Markov decision processes, the value function $V^\pi$ associated with a deterministic policy $\pi : \mathcal{X} \to \mathcal{A}$ is defined as $V^\pi(x) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| x_0 = x\right]$ for all states $x \in \mathcal{X}$. Using this notation, give the definition of an optimal policy $\pi^*$ and the optimal value function $V^*$.

**Question 9:** [1 point] The Value Iteration algorithm for discounted Markov decision processes iteratively computes a sequence of value functions $V_1, V_2, \ldots, V_k$. In a given iteration $k$ of the algorithm, what is the relation between $V_k$ and $V_{k+1}$?

**Question 10:** [2 points] Describe the TD(0) algorithm for policy evaluation in discounted Markov decision processes. If the algorithm is luckily initialized with the true value function $\widehat{V}_0 = V^\pi$, what is the expectation of $\widehat{V}_1$?