General rules:

- The Midterm will take place on Tuesday, February 26, in class (4-5:20pm).
- You are responsible for all material up to (and including) SVMs (but excluding gradient descent)
- No aids allowed (no notes, no books, no calculators), just bring a pen

Key concepts

The following is a (not complete) list of key concepts you should be familiar with:

- Induction versus deduction
- Polynomial curve fitting
- General framework of machine learning (feature space, class of predictors, loss function, empirical losses (training loss, test loss), true loss
- ERM and RLM
- Overfitting and how to prevent it
- Bayesian reasoning (prior distribution, likelihoods, MLE, MAP)
- Regression and classification
- Loss functions used in regression and classification (be able to give examples)
- Linear regression (design matrix, quadratic loss, non-linear basis functions, ERM and RLM solution, interpretation as MLE and MAP)
- Bayes predictor and Bayes risk (in regression with quadratic loss, in classification with 0/1-loss)
- Linear classifiers (different methods to learn them: perceptron, LDA, hard and soft SVM, logistic regression)
- Data generation with Gaussian class-conditionals (Bayes optimal, maximum likelihood estimation of parameters)
- Convexity and surrogate loss functions

Practice questions

1. For both classification and regression, describe situations where the empirical loss of an ERM predictor is very small but the true risk is large. How can we prevent this from happening?

2. Describe a dataset on which LDA will not find an ERM solution. Explain why.

3. Give an example of a data set $D = ((x_1, t_1), \ldots (x_{10}, t_{10}))$ on which the Perceptron algorithm will stop after one update.

4. Prove or refute: If $f, g : \mathbb{R}^d \to \mathbb{R}$ are convex functions, then their product $f \cdot g : \mathbb{R}^d \to \mathbb{R}$ defined as $f \cdot g(x) = f(x) \cdot g(x)$ is convex.

5. We have shown in class that, if the class conditional densities $p(x|1)$ and $p(x|0)$ are both Gaussian distributions, then the Bayes classifier is a linear classifier. What happens if the class conditionals are each uniform over some ball in $\mathbb{R}^d$? Say both these balls have the same radius. How does it depend on the class priors $p(1)$ and $p(0)$?

6. A bag contains four dice: a 4-sided die, two 8-sided dice and a 12-sided die. A person chooses a die from the bag at random and rolls it. The number of faces that the selected die has is denoted by the random variable S and the number rolled by the random variable X.

   (a) Derive the probability distribution of the number rolled with the selected die. Use your expression for $p(X)$ to determine the probability that the first number rolled is 3. That is, what is $p(X = 3)$?

   (b) Given that the roll was a 3, what is the posterior probability that the selected die had 12 sides? That is, what is $p(S = 12|X = 3)$?

   (c) Suppose the person rolls the same die a second time. Denote the random variable of this second roll by Y . If the second roll comes up a 7, how does this change the posterior probability that the selected die has 12 sides? Specifically, what is $p(S = 12|X = 3, Y = 7)$?

7. Is ERM with respect to the square loss (in linear regression) a convex optimization problem? Explain why or why not!

8. Explain the concept of a surrogate loss function. Why do we use surrogate loss functions? Which properties should they have? Give an example of such a loss function.

9. For an RLM approach to linear regression, we aim to minimize $\mathcal{L}_D^{square}(y_{\mathbf{w}}) + \lambda\|\mathbf{w}\|^2$, where $\mathcal{L}_D^{square}(y_{\mathbf{w}})$ denotes the empirical (square) loss of the training dataset D. Explain the role of the parameter λ here. What happens for very large values of λ, what happens for very small values of λ? How can you determine what is a good value for λ?

10. Similarly in SVM, we minimize $\mathcal{L}_D^{hinge}(y_{\mathbf{w}}) + \lambda\|\mathbf{w}\|^2$. Again, discuss the role of λ.