Selected problems in Machine Learning – warm-up test

1. Plot the following functions: (a) $f(x) = x\exp(x)$, (b) $f(x) = \ln\frac{x^2-1}{x^2+1}$, (c) $f(x) = \left|\frac{x-1}{1-2x}\right|$, (d) $f(x) = \sqrt{1-\exp(-x^2)}$
2. Define convex function.
3. Define convex region in R^2.
4. What is gradient?
5. What is Hessian matrix?
6. Find a growing function which maps R to $< 0, 1 >$.
7. Given joint distribution $p(A, B)$, express $p(A)$ and $p(A|B)$.
8. Explain “curse of dimensionality”.
9. Explain the main difference between the frequentist and Bayesian interpretation of probability.
10. Derive Bayes’ theorem.
11. Define independence (independent random variables X and Y).
12. Define conditional independence (variable X independent of Y given Z).
13. What does it means that a collection of random variables (e.g. a sequence) is i.i.d. (independent and identically distributed).
14. What is the relation between a probability density function and associated cumulative distribution function?
15. Explain the difference between the terms probability and likelihood.
16. Why we use probabilistic methods?
17. Where does uncertainty in NLP tasks come from?
18. What is correlation?
19. What is variance?
20. What is covariance matrix?
21. Let’s suppose that the sequence (1, 3, 4, 4, 8) is drawn from $\mathcal{N}(\mu, \sigma^{\in})$. What are the values of μ and σ^2 (according to Maximum Likelihook)?
22. Plot probability densities $p(r)$ and $p(\phi)$ which result from transforming 2D Gaussian distribution centered in the origin with unit covariance matrix into polar coordinates.
23. What can you say about a multidimensional Gaussian with covariance matrix having zeros everywhere outside its diagonal.
24. Why is Gaussian distribution so special?
25. Which types of random variables cannot be modeled by Gaussian distributions?
26. Could you sketch a histogram for lenght distribution of dependency relations? (distinguish orientation)
27. Could you sketch a histogram for lenght distribution of anaphoric relations? (distinguish orientation)
28. What are Monte Carlo methods used for?
29. Explain how Gibbs sampling works.
30. For which distributions you can *not* use Gibbs sampler?
31. If you have a generator of numbers from the uniform distribution on $[0, 1]$, how would you generate samples with probability density $p(X)$, $x \in R$.
32. How can you generate samples from a uniform distribution in 3D unit ball?
33. How can you generate samples from 1D Gaussian distribution?
34. Suppose you have a sampler of x-y pairs from a joint distribution $p(x, y)$. How can you generate samples from the asso-

ciated marginal distribution $p(x)$.

35. What is entropy?
36. Which distribution on discrete values has the lowest entropy?
37. Which distributution on continues values has the lowest entropy?
38. What is mutual information?
39. What is KL divergence? Why it is not a distance (in the mathematical sense)?
40. Explain the difference between generative models and discriminative models.
41. Explain the difference between classification and regression.
42. What is separation boundary (in classifiers)?
43. What does it mean that two sets of points are linearly separable.
44. Explain how entropy maximization can be used in classification.
45. Explain how entropy minimization can be used in classification.
46. Explain how Naive Bayes classifier works.
47. What is the main idea behind SVM kernels?
48. Plot the sigmoid function used in logistic regression ($f(z) = \frac{1}{1+e^{-z}}$).
49. Assume there are two classes of points in 2D in the training data: $A = (0, 0), (1, 1)$ and $B = (3, 1), (2, 3), (2, 4)$. Find some separation hyperplane and write its equation. Could you sketch separation boundaries that would be found by (a) SVM, (b) perceptron, (c) 1-NN?
50. Design feature transformation functions which make the following sets linearly separable: (a) $A = (2, 0)$ and $B = (1, 0), (3, 0), (2, 1), (2, -1)$, (b) $A = (0, 0), (2, 2)$ and $B = (1, 1), (3, 3)$.
51. Explain how k-means clustering works.
52. Explain how EM works.
53. What is cost function (loss function)?
54. What is bias-variance trade-off.
55. Illustrate underfit/overfit problems in regression by fitting a polynomial function to a sequence of points in 2D.
56. Illustrate underfit/overfit problems in classification by modeling two (partially overlapping) classes of points in 2D.
57. Plot a typical dependence of training and test error rates (vertical axis) on training data size (horizontal axis).
58. Plot a typical dependence of training and test error rates (vertical axis) on model complexity (horizontal axis).
59. What problem is typically signalled by test error being much higher that training error?
60. What problem is typically signalled by test and training errors being stabilized after a limited portion of training data?
61. What is regularization used for?
62. Explain the difference between parametric and non-parametric methods.
63. Name some quantities that can be used for feature selection.