This is a test final exam for Machine Learning II. The final exam will be multiple choice only, with multiple correct answers. This test exam demonstrates the degree of difficulty and covers the great majority of topics that are needed.

1. Keep the Balance

   Let X denote data, $X = (x_1, \ldots, x_n)$, with samples $x_i \in \mathbb{R}$, $i = 1, \ldots, n$, $n \in \mathbb{N}$. Which linear transformation $T : x \rightarrow x'$ guarantees zero mean and unit variance for X? Provide the explicit mathematical form of T, i.e., $x_i \rightarrow x'_i = \frac{x_i - ?}{?}$. T is a function of all data samples **and** the number n. Hint: Define and use both the sample mean and the sample variance to construct T. No proof needed for the correctness of your given result for T.

   Solution: $T : x_i \rightarrow \frac{x_i - \overline{x}}{\sigma}$ with $\overline{x} = \frac{1}{n} \sum x_i$ and $\sigma^2 = \frac{1}{n-1} \sum (x_i - \overline{x})^2$.

2. Fantastic and Elastic Net Terms

   (a) Use the notation w_i, $i = 0, \ldots M$, for the weights/slopes. What are the two regularization terms in Elastic Net,
   (i) using explicit summation with respect to the w_i's ?

   Solution:

   $$+\lambda_1 \sum_{j=0}^{M} |w_j| + \lambda_2 \sum_{j=0}^{M} w_j^2$$

   (ii) expressing one of the terms in terms of the L1 norm ($||\cdot||_1$) and one term in terms of the L2 norm ($||\cdot||_2$), i.e. without using explicit summation?

   Solution:

   $$+\lambda_1 ||w||_1 + \lambda_2 ||w||_2^2$$

   (b) What does regularization do? Answer simply *yes* or *no*, no explanation needed:
   (i) *making the regression slopes much larger, to avoid getting stuck in gradient descent*
   (ii) *increase both variance and bias*
   (iii) *increase only variance*
   (iv) *increase only bias*

   Solution/Answers: all NO, except (iv): YES

3. Prasanta Chandra Mahalanobis from the Distance

   Let

   $$d(\mathbf{x}, \mathbf{y}) = \left[(\mathbf{x} - \mathbf{y})^T S^{-1} (\mathbf{x} - \mathbf{y}) \right]^{1/2}$$

be the generalized interpoint distance (Mahalanobis distance), for two sample points $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, with S being the covariance matrix for the distribution of the data $D \subset \mathbb{R}^n$.

T indicates the transposed vector. S^{-1} is the inverse of S.

Show for $S = \mathbb{I}$ (representing unit variance and zero off-diagonals; $\mathbb{I}$ is the unit matrix) that $d(\mathbf{x}, \mathbf{y})$ reduces to the ordinary Euclidean (L2) distance.

Solution:

$$d(\mathbf{x}, \mathbf{y}) = \left[(\mathbf{x} - \mathbf{y})^T S^{-1} (\mathbf{x} - \mathbf{y})\right]^{1/2} =$$

$$\left[(\mathbf{x} - \mathbf{y})^T S (\mathbf{x} - \mathbf{y})\right]^{1/2} = \left[(\mathbf{x} - \mathbf{y})^2\right]^{1/2} =$$

$$||\mathbf{x} - \mathbf{y}||_2$$

4. Principal Component Analysis (PCA)

The PCA analysis of a given dataset $X \subset \mathbb{R}^{n \times m}$, with $m = 5$ is the feature dimension, and $n = 1234$ samples, yields the eigenvalues $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5 = 5, 4, 3, 2, 1$, to the eigenvectors v_1, v_2, v_3, v_4, v_5.

(a) Assume that $||v_i||_2 = 1$, for $1 \leq i \leq m$, and that the dot product $v_i \cdot v_j$ is defined as the sum over the products of the vector components. Which values of $v_i \cdot v_j$ are taken for *each* of the possible dot product combinations for v_i and v_j (cases $i = j$ included)?

Solution: PCA components show orthogonality.

(b) What is the **minimal** reduced dataset dimension, given it is required that at least 4/5 of the "variance is explained" in terms of eigenvalue importance/eigenvalue weights? Which set of tuples (eigenvector, eigenvalue) you would choose for this case?

Solution: 4:5 = (5+4+3):(5+4+3+2+1)=12/15. Tupels: (5,v_1), (4,v_2), (3,v_3).

5. Andrei Andreyevich Markov in Chains

Fig. 1(a,b,c,d) shows 4 Markov processes. Answer the questions, no explanation needed:
(i) Which processes (a,b,c,d) exhibit the normalization property?
(ii) Which processes that show normalization are ergodic?
(iii) For the processes that show normalization but no ergodicity, characterize in your words the stationary distribution!

Solutions:
(i): a, c, d

(ii): a
(iii): c: switching between 2 and 3, d: final state 1

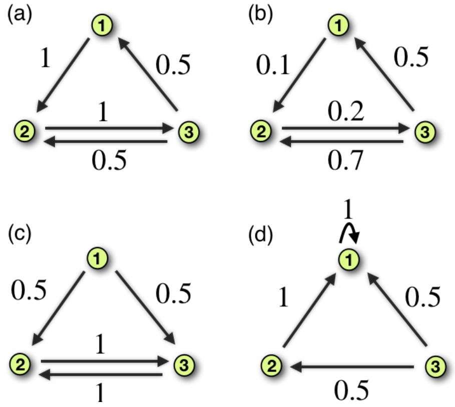

Figure 1: 4 Markov processes (a-d) of 3 states with transition probabilities.

6. MCMC Sampling

   Which of the following statements is correct?

   (i) Hamiltonian Monte Carlo sampler is a Gibbs sampler
   (ii) Gibbs sampler suffers heavily from step size convergence problems
   (iii) Metropolis is more general than Metropolis-Hastings
   (iv) NUTS is an elliptic sampler that slices through marginal symplectic surrogates
   answer: all incorrect.

7. maybe: Python

   What does the code do? 4 choices, more than 1 might be correct. Code simple. No preview given.