(only FALSE ones are indicated or when there has been some ambiguity.)

1. **Complexity control and all that jazz**.

   - ❏ The VC dimension of a hypothesis class tells us how capable its members are to fit data.
   - ❏ In machine learning it is necessary to have classifiers with zero training error. **FALSE, zero training error may lead to overfitting for example**
   - ❏ Regularization is a framework in which poor goodness-of-fit can be compensated with complexity. **FALSE, here when I wrote "complexity" I meant high complexity, but I have also accepted TRUE answers if explained that low complexity was interpreted.**
   - ❏ The VC dimension can only be infinite for models with infinite parameters. **FALSE, in fact in class we saw an example of a function class with 1 parameter and infinite VC dimension**
   - ❏ Test data is typically used to estimate generalization error. also validation data
   - ❏ Validation data is typically used to perform model selection.
   - ❏ The VC dimension of a function strongly affects its variance during learning. **FALSE, "VC dimension of a function" makes no sense as VC dimension applies to a family of functions or hypothesis class.**
   - ❏ The VC dimension is related to the number of parameters of functions in a hypothesis class. **FALSE, not necessarily (example with 1 parameter and infinite dimension for example.)**
   - ❏ Empirical risk, the opposite of training error, serves as an approximation to the true risk. **FALSE, empirical risk is not the opposite of training error.**
   - ❏ In order to prove that VC dimension is strictly smaller than n, we have to show that every set of cardinality n cannot be shattered.

2. **Bayesian classifiers**.

   - ❏ Linear models are called linear because they are linear in their parameters.
   - ❏ By definition, Bayes rule chooses the class with highest posterior probability.
   - ❏ Bayesian classifiers lead to linear decision boundaries. **FALSE, not always, for example QDA leads to quadratic decision boundaries.**
   - ❏ The bayesian classifier is optimal in the sense that it has the smallest generalization error among all classifiers. aru ley pahila data ko distribution assume garcha, esle data herera distribution determine garcha ani prediction garcha
   - ❏ The Mahalanobis distance is a generalized form of Euclidean distance that takes into account correlations among variables.
   - ❏ If the class-conditional distributions in a classification problem are normal, then linear functions can achieve the best possible generalization error. **FALSE, in QDA we assume normal class-conditional distributions but optimal classifier is not linear in general.** linear cha bhane optimal didaina, complexity add garda matra better huncha
   - ❏ When using QDA or LDA in a practical problem with finite data that is normally distributed, there is no need to estimate generalization error because we know that these methods are optimal. **FALSE, with finite data we still need to estimate the parameters of the gaussians and therefore we still need to check quality of the trained models.**
   - ❏ Laplace correction is particularly necessary in Naive Bayes classification if the dimensionality of the input data is very large. **My particular answer here would be TRUE, however I have accepted some FALSE ones that have argued that the problem is with 0 counts. The reasoning behind my TRUE is that the "probabiltiy" that 0 counts happen grows exponentially with the dimensionality of the input, and therefore we need to deal with this.**
   - ❏ The VC dimension of the k-NN classifier is infinite.

- ❑ Tuning the "k" in k-NN trades off bias and variance.

3. **Maximum Likelihood and GLMs**.

   - ❑ The maximum likelihood method is a general method based on the optimization of the likelihood function which is a function determined by a data sample that maps parameter values to non-negative real values.
   - ❑ The likelihood function is a probability distribution over all possible parameter values for finite sets of data. **FALSE, if you add (or integrate) over all parameters values the results does not have to be 1 (for example).**
   - ❑ GLMs are typically trained using the maximum likelihood method which involves solving some form of linear optimization problem. **FALSE, optimizing likelihood typically involves non-linear optimization techniques.**
   - ❑ Logistic regression is considered a (generalized) linear model because the classes' posteriors are linear functions of the predictors. **FALSE, the log of the odds is a linear function of the predictors, not the posteriors.**
   - ❑ The Iterated Reweighted Least Squares algorithm is used to find the maximum likelihood optimum in logistic regression.
   - ❑ The decision boundary of a logistic regression model has a sigmoidal shape. **FALSE, the decision boundary is linear.** decision boundary is always linear. Sigmoid function is applied as activation function sometimes
   - ❑ The likelihood function to be optimized when modelling with GLM is determined by the target distribution $P(T|\mathbf{X};\theta)$ assumed in the model.
   - ❑ Zero bias and variance of a linear regression model implies zero training error. **FALSE, we still have noise in the training data.**
   - ❑ Fitting non-linear models requires non-linear optimization techniques. **FALSE, this is not always so, e.g. in SVMs we still use linear techniques to fit nonlinear functions by making use of the kernel trick.**
   - ❑ The "generalized" part of GLM is due to their use of link functions.

4. **Neural networks**.

   - ❑ MLPs generalize GLMs
   - ❑ Once trained, MLP networks are deterministic while RBF networks are not. **FALSE, both MLPs and RBFs are deterministic.**
   - ❑ The only differences between MLP and RBF networks are the number of hidden layers they allow and their training mechanisms. **FALSE, they use different "types" of nodes as well.**
   - ❑ Since GLMs are particular cases of MLPs, the backpropagation algorithm is enough to train a logistic model algorithm. **FALSE, the backpropagation algorithm only computes gradients, we need to use these gradients in some way to update parameters.**
   - ❑ Any cost function, as long as it is differentiable, can be used in the backpropagation algorithm to find suitable weights.
   - ❑ Computing the gradients during MLP training requires an iterative procedure which may not always converge. **FALSE, we can use backpropagation which always finishes, all it does is one forward and one backward step through the layers of the network.**
   - ❑ Backpropagation is a non-deterministic procedure which, depending on how initial values are chosen may give different results. **FALSE, backpropagation is totally deterministic.**
   - ❑ Gradient descent is a non-deterministic procedure which, depending on how initial values are chosen may give different results.
   - ❑ The number of neurons in hidden layers may have a large impact on the quality of the predictions of a neural network.

- ❑ RBF neural networks can be trained using backpropagation. **FALSE, the first layer is "trained" using some sort of clustering, and the second layer can be computed using some sort of gradient descent method, for example.**

5. **Kernels and SVMs**.

- ❑ Adding two kernels k_1 and k_2 with corresponding transformations ϕ_1 and ϕ_2 yields a new kernel k_3 which corresponds to the data transformation $\phi_3 : \mathbf{x} \mapsto \phi_1(\mathbf{x}) + \phi_2(\mathbf{x})$. **FALSE, adding two kernels corresponds to the concatenation of transformations.**
- ❑ Multipying a kernel k by a strictly positive scalar yields a new kernel k' whose corresponding transformation is of dimensionality at most that of k's transformation.
- ❑ Since adding kernels and multiplying them by scalars yield new kernels, any polynomial on a given kernel is also a kernel. **FALSE, for example -k is a polynomial on k (k being a kernel) and it is not a kernel.**
- ❑ There is no danger of overfitting in SVMs because seeking large margins prevents it. **FALSE, using some kernels effectively adds complexity to the models and therefore there is danger of overfitting.**
- ❑ Increasing the value of C (all else being equal) in an SVM may decrease training error.
- ❑ Kernel matrices are always positive semi-definite because they correspond to inner products in some Hilbert space. **My answer here is TRUE but I also accepted FALSE answers that pointed out the existence of CPSD kernels.**
- ❑ SVMs produce linear classifiers with high margin if we chose an appropriate kernel. **FALSE, for some data this is not possible**
- ❑ An algorithm can be kernelized if its solution can be expressed as a linear combination of input vectors.
- ❑ Kernels are defined as functions of two input vectors in Euclidean space whose corresponding kernel matrices are positive semi-definite. **FALSE, kernels do not necessarily need to operate on vectors in Euclidean space.**
- ❑ Linear SVMs tend to generalize better because simpler models have lower VC dimension. **FALSE, depends on the data.**

6. **Miscellaneous**.

- ❑ Unsupervised learning is about making predictions on unseen future examples. **FALSE, in unsupervised learning there is no notion of making predictions.**
- ❑ High training and test error is an indicator of high bias in a model.
- ❑ High training error and low test error is an indicator of overfitting in a model. **FALSE, it would be low training and high test error.**
- ❑ Random forests are typically better predictors than individual decision trees due to the high variance of individual trees, which random forests reduce.
- ❑ The E-M algorithm is an optimization algorithm that maximizes some likelihood function.
- ❑ The E-M algorithm is guaranteed to find an optimum solution if the input data is distributed according to a mixture of gaussians. **FALSE, the E-M algorithm only finds local optima.**
- ❑ If all elements in a mixture of gaussians have equal covariance matrices, then E-M behaves like k-means. **FALSE, the covariance matrices also need to be proportional to the diagonal matrix**
- ❑ Clustering makes little sense in practice since we have no way of knowing whether the result is fully truthful. **FALSE, clustering can be very useful in practice and it is very much used.**
- ❑ Classification is always easier than regression or clustering. **FALSE, it depends on the particular problem (data).**
- ❑ Clustering can be hard because in most cases we have no gold standard.