**知识点总结：**

**1.1**

The course starts with the concept of statistical machine learning, which aims to extract features from data sets and use them for prediction. Basically, statistical machine learning can be divided into supervised learning and unsupervised learning. In addition, there are semi-supervised learning, active learning, multi-task learning, etc., which will not be covered in this course. First of all, the course reviews the nature of the multivariate normal distribution. It is necessary to understand that the linear combination of the multivariate normal distribution and the conditional distribution is still a normal distribution. Then introduced frequency inference and Bayesian inference, parametric model and non-parametric model.

KNN is an important classification and regression algorithm. The value of k and the measurement of distance are very important, which will affect the under-fitting and over-fitting of the model. As the dimension increases, the generalization ability of KNN will generally decrease, so PCA is introduced for dimension reduction.

**1.2**

As a typical unsupervised learning algorithm, the performance of KNN will decrease significantly when the dimensionality increases: we need a longer scale to cover a certain proportion of the space. Therefore Principal Component Analysis (PCA) is introduced for dimension reduction, and the mathematical derivation is completed in class. It should be noted that mathematical derivation needs to use some common conclusions: symmetric matrices on the real field can be diagonal by orthogonal matrix, the definition of sample covariance matrix, and Lagrangian multiplier method. Then we discuss the loss function and risk (that is, the expectation of the loss function, but because the joint distribution is unknown, it cannot be accurately calculated). Thus, the data set can be divided into a training set and a test set. The loss on the training set The mean of the function is called the experience loss, and the experience loss is not an unbiased estimate of risk. For the test set, it can be proved that the mean value of the loss function on the test set can be used as an unbiased estimate of risk.

Regularization is an important method of model selection, which aims to reduce model complexity and effectively prevent overfitting. Ridge regression is a prime example. Naive Bayes estimation is also an important algorithm for classification, in which the independence between data features is assumed. The discriminant function can be used as a basis for classification judgment.

**1.3**

The course focuses on the decomposition formula of expected loss. Expected loss (that is, risk) can be decomposed into three parts: noise, variance and bias. First of all, students need to master the double expectation formula, in the case of fixed X, consider the variance of Y, and then find the expectation of X, which is noise. Variance refers to the robustness of a given algorithm, that is, after a given hypothesis space H and model f, for different training sets D, the difference of the model is evaluated. Bias refers to the average meaning of f (here the average refers to the average of the training set D), the difference with respect to the theoretical optimal model f\*. Generally speaking, it is difficult to reduce variance and bias at the same time, so the theory of balance between variance and bias is introduced.

Finally, the linear discriminant model (LDA) is introduced. LDA generates a model based on the maximization of posterior probability. Using Bayesian theory, the posterior probability can be calculated and the discriminant function can be obtained. Students are required to have a good grasp of the posterior concept of the Bayesian school, and be able to deduce the posterior density of the parameters and the posterior conditions of Y after the given sample set D from the prior density of the parameters and the conditional density of Y under the given parameters density.

**1.4**

The course first distinguishes between discriminative models and generative models in machine learning. In General, A discriminative model ‌models the decision boundary between the classes. A Generative Model ‌explicitly models the actual distribution of each class. A Generative Model ‌learns the joint probability distribution p(x,y). It predicts the conditional probability with the help of Bayes Theorem. A discriminative model ‌learns the conditional probability distribution p(y|x). Both of these models were generally used in supervised learning problems.

Then the course introduces logistic regression. The log odds ratio is assumed to be a hyperplane, and our goal is to estimate the parameter values in the hyperplane. Support vector machines also use hyperplanes for classification algorithms. In simple terms, support vector machines can be divided into linear separable support vector machines and linear support vector machines. In the latter, slack variables are introduced, and their essence is Quadratic convex optimization problem.

**1.5**

The course begins with a discussion on the mathematical derivation of SVM. The course introduces the concepts of geometric interval and function interval respectively, and then introduces optimization methods based on the idea of minimizing the function interval. The generalized Lagrangian function can be introduced to transform the original problem into a minimax problem. We prove that the solution of its dual form is not greater than the solution of the original problem. Then, in some special cases, the solutions of the original problem and the dual problem are equal. At this time, the KKT condition can be used to derive. Using the above optimization method, we can give a closed form solution of the linearly separable support vector machine.

The introduction of the kernel function allows us to effectively solve the classification problem of a class of curve boundaries. An implicit mapping maps the input space to the feature space (Hilbert space), and then implements supervised learning and Unsupervised learning algorithm. The Repoducing Kernel Hilbert Space (RKHS) is introduced in detail, and the Mercer theorem: The necessary and sufficient condition for a symmetric function to be a Mercer kernel is that its Gram matrix is positive semi-definite.