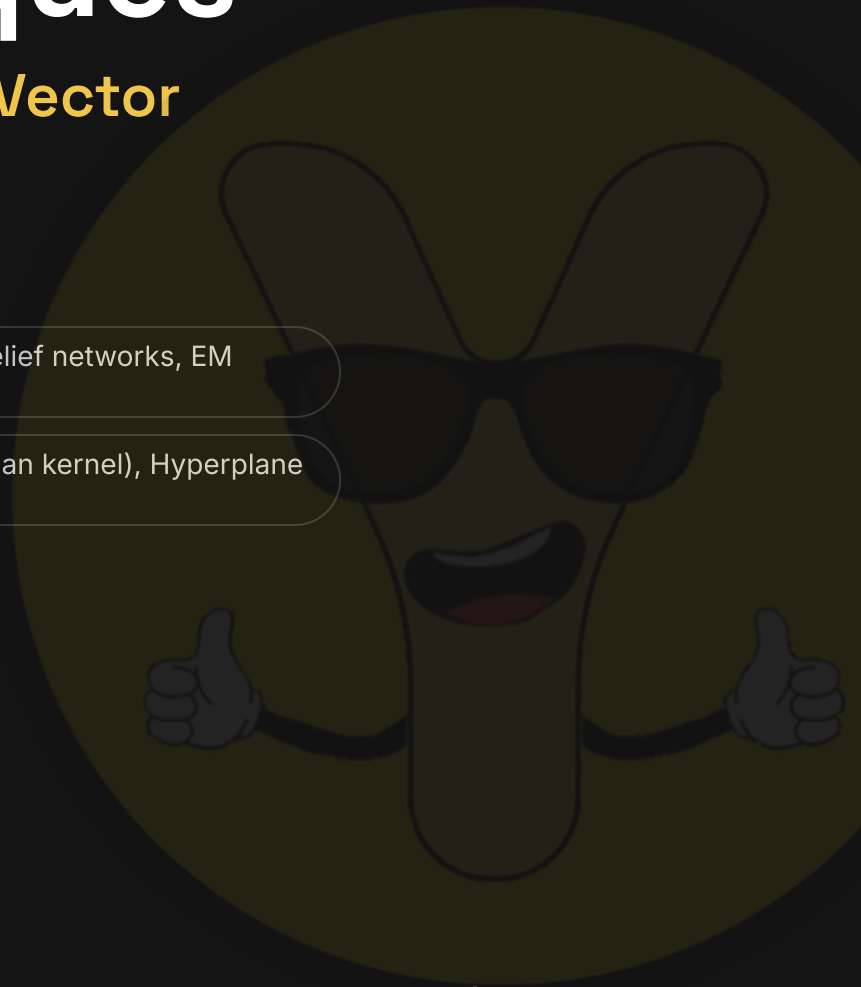
AKTU · 3RD YEAR · SEMESTER 6 · BCDS062

Machine Learning Techniques

Unit 2 — Regression, Bayesian Learning and Support Vector Machines

- Regression: Linear Regression and Logistic Regression
- Bayesian Learning: Bayes theorem, Concept learning, Bayes Optimal Classifier, Naïve Bayes classifier, Bayesian belief networks, EM algorithm
- Support Vector Machine: Introduction, Types of support vector kernel (Linear kernel, polynomial kernel and Gaussian kernel), Hyperplane (Decision surface), Properties of SVM, Issues in SVM



DECODED BY

Chirag Jain

Founder, Ycotes · Gold Medalist in Artificial Intelligence, IIT Madras · taught 10,000+ students

theycotes.com

Unit 2 Overview: Regression, Bayesian Learning, and SVM

CORE PARADIGMS

This unit covers parametric regression models for continuous and categorical targets, probabilistic Bayesian inference algorithms, and margin-maximizing Support Vector Machines.

- Linear and Logistic Regression for prediction and classification.
- Bayesian learning foundations, Naive Bayes, and the EM algorithm.
- Support Vector Machines, kernel tricks, and optimal hyperplanes.

Linear Regression

Linear regression models the continuous relationship between a dependent target and one or more independent features.

- Univariate hypothesis function: $h_w(x) = w_0 + w_1 x$.
- Multivariate hypothesis function: $h_w(x) = w_0 + w_1 x_1 + \cdots + w_n x_n = W^T X$.
- Cost function (Mean Squared Error): $J(W) = \frac{1}{2m} \sum_{i=1}^{m} (h_w(x^{(i)}) - y^{(i)})^2$.
- Objective: Minimize $J(W)$ using Gradient Descent update rules.
- Gradient descent step: $w_j := w_j - \alpha \frac{\partial}{\partial w_j} J(W)$.

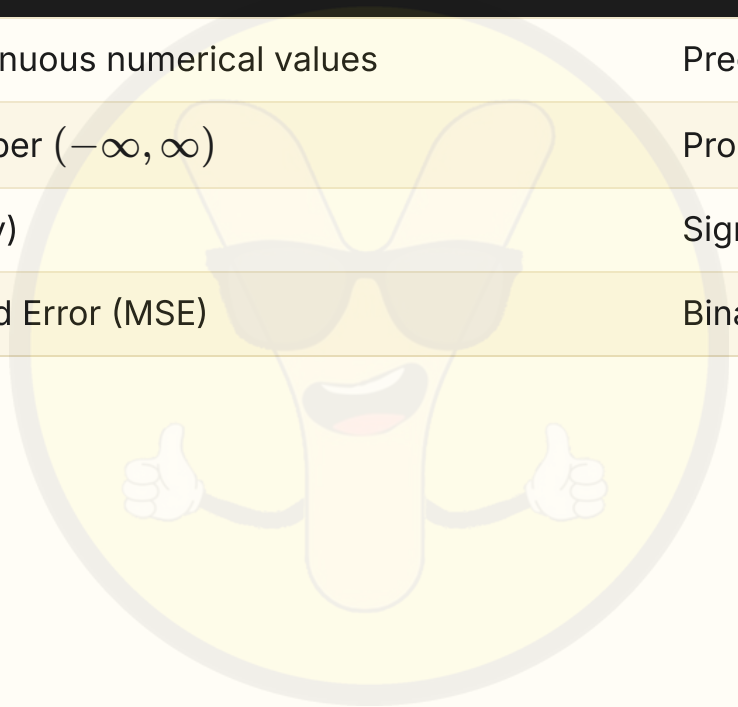
Logistic Regression

Logistic regression is used for binary classification by mapping predicted values to probabilities between 0 and 1.

- Uses the Sigmoid (Logistic) function: $g(z) = \frac{1}{1+e^{-z}}$.
- Hypothesis function: $h_w(x) = g(W^T X) = \frac{1}{1+e^{-W^T X}}$.
- Decision boundary: Predict 1 if $h_w(x) \geq 0.5$ (i.e., $W^T X \geq 0$), else predict 0.
- Cost function (Log Loss / Binary Cross-Entropy): $J(W) = -\frac{1}{m} \sum [y^{(i)} \log(h_w(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_w(x^{(i)}))]$.
- Gradient descent update has the exact same algebraic form as linear regression, but with the sigmoid hypothesis.

Linear vs. Logistic Regression

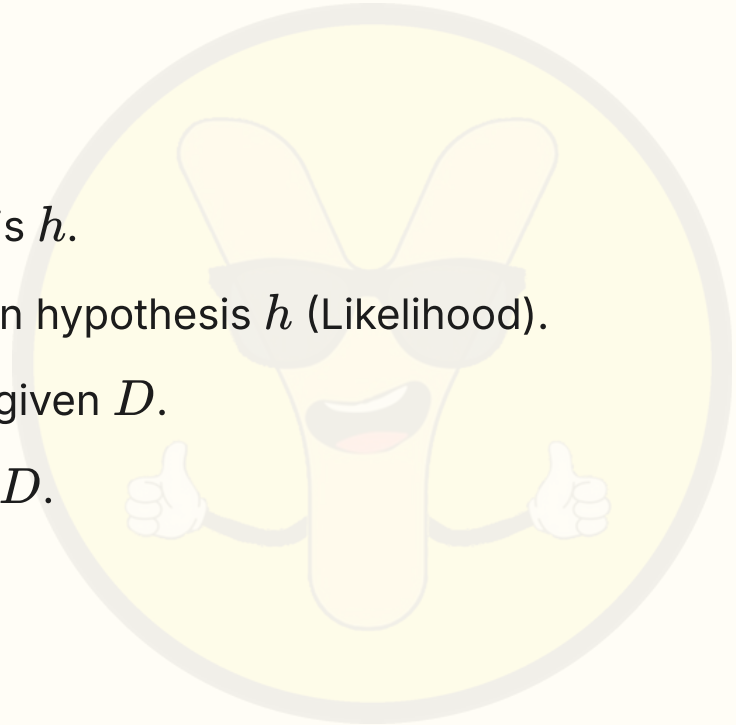
Feature	Linear Regression	Logistic Regression
Task	Predicts continuous numerical values	Predicts categorical binary classes
Output Range	Any real number $(-\infty, \infty)$	Probabilities between $[0, 1]$
Activation Function	None (Identity)	Sigmoid Function
Cost Function	Mean Squared Error (MSE)	Binary Cross-Entropy (Log Loss)



Bayes Theorem and Fundamentals

Bayes Theorem provides a mathematical foundation for calculating conditional probabilities given prior knowledge.

- Formula: $P(h|D) = \frac{P(D|h)P(h)}{P(D)}$.
- $P(h)$: Prior probability of hypothesis h.
- $P(D|h)$: Probability of data D given hypothesis h (Likelihood).
- $P(h|D)$: Posterior probability of h given D.
- $P(D)$: Marginal probability of data D.



Concept Learning and Bayes Optimal Classifier

Concept learning framed probabilistically seeks the most probable hypothesis from training data.

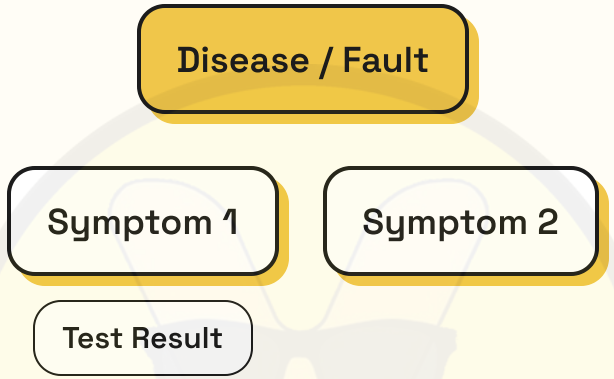
- Maximum A Posteriori (MAP) hypothesis: $h_{MAP} = \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} P(D|h)P(h)$.
- Maximum Likelihood Estimation (MLE): If all priors $P(h)$ are equal, $h_{MLE} = \arg\max_{h \in H} P(D|h)$.
- Bayes Optimal Classifier: Calculates the most probable output value given new data, combining predictions of all hypotheses weighted by their posterior probabilities.
- Formula: $P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$.
- Classification rule: $\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$.

Naive Bayes Classifier

The Naive Bayes classifier simplifies learning by assuming conditional independence among feature variables given the class label.

- Classification rule: $v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i|v_j)$.
- Avoids zero-frequency problem using Laplace (additive) smoothing: $\hat{P}(a_i|v_j) = \frac{n_c+\alpha}{n+\alpha\cdot|V|}$.
- Advantages: Highly efficient, requires small training data, handles high-dimensional data well (text classification).
- Disadvantage: Conditional independence assumption is often violated in real-world correlated features.

Bayesian Belief Networks Architecture



Directed Acyclic Graph representing conditional dependencies among random variables in a Bayesian Belief Network.

Expectation-Maximization (EM) Algorithm

The EM algorithm finds maximum likelihood estimates of parameters in probabilistic models with latent (unobserved) variables.

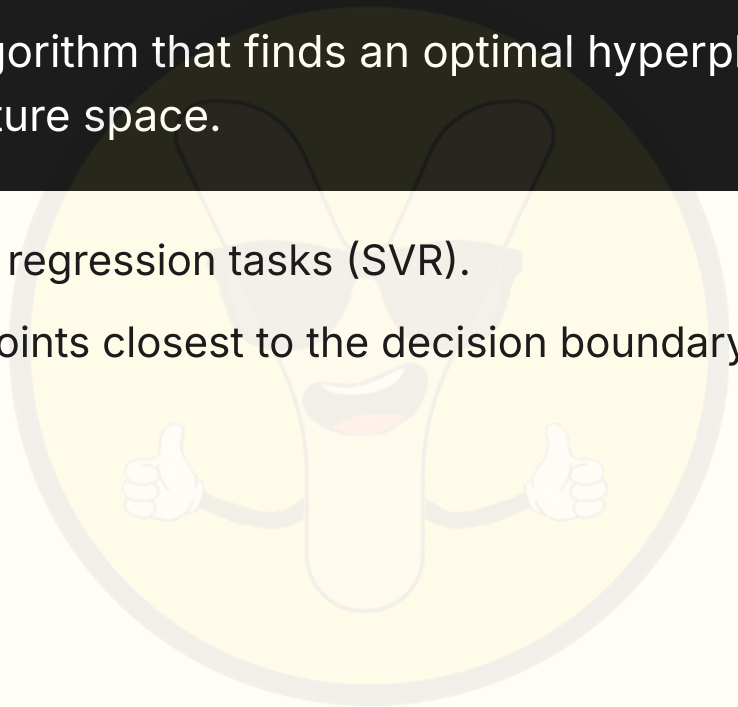
- Iterative two-step optimization process ensuring convergence to a local optimum.
- E-step (Expectation): Calculate the expected value of the log-likelihood function with respect to the conditional distribution of latent variables.
- M-step (Maximization): Update parameters to maximize the expected log-likelihood found in the E-step.
- Common applications: Gaussian Mixture Models (GMM), clustering with missing data.

Support Vector Machine: Introduction

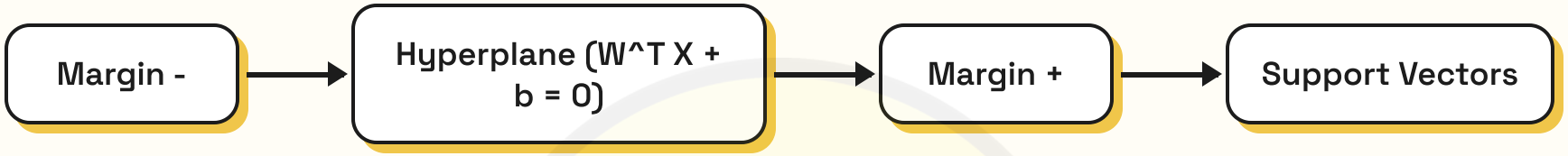
SVM

A supervised machine learning algorithm that finds an optimal hyperplane maximizing the margin between different classes in a feature space.

- Operates on both classification and regression tasks (SVR).
- Focuses only on support vectors (points closest to the decision boundary) rather than all training data.



SVM Hyperplane and Margin Geometry



Linear SVM separating two classes with maximum margin $\frac{2}{||W||}$.

Hyperplane and Decision Surface

The decision surface in SVM is a hyperplane separating classes in an N-dimensional space.

- Linear equation: $W^T X + b = 0$, where W is the weight vector and b is bias.
- Decision rules: $W^T X + b \geq +1$ for class +1, and $W^T X + b \leq -1$ for class -1.
- Margin width: Defined geometrically as $\frac{2}{||W||}$.
- Optimization objective: Minimize $\frac{1}{2}||W||^2$ subject to constraints $y^{(i)}(W^T x^{(i)} + b) \geq 1$.

Types of Support Vector Kernels

Kernels transform non-linearly separable input data into a higher-dimensional space where a linear hyperplane can separate them.

- Linear Kernel: $K(x, x') = x^T x'$. Used for linearly separable data.
- Polynomial Kernel: $K(x, x') = (gamma \cdot x^T x' + c)^d$. Captures feature interactions up degree d.
- Gaussian Radial Basis Function (RBF) Kernel: $K(x, x') = \exp(-\gamma ||x - x'||^2)$. Maps data to infinite-dimensional space.
- Kernel Trick avoids explicit high-dimensional coordinate calculations, computing inner products directly.

Properties and Issues of SVM

Understanding structural risk minimization and practical limitations of Support Vector Machines.

- Properties: Excellent generalization due to maximum margin, effective in high dimensions, robust against overfitting with proper regularization.
- Issues: Computationally intensive on large datasets $O(n^3)$ training time.
- Issues: Highly sensitive to noisy data and outliers.
- Issues: Selecting the optimal kernel function and hyperparameters (like C and γ) requires exhaustive cross-validation.

Exam corner

- **Q1** Derive the cost function and gradient update equations for Logistic Regression.
- **Q2** State and prove Bayes Theorem. Differentiate between MAP and MLE estimators.
- **Q3** Explain the Naive Bayes classifier with an example and discuss Laplace smoothing.
- **Q4** What is the EM algorithm? Explain its Expectation and Maximization steps.
- **Q5** Define Support Vector Machine. Explain the role of hyperplanes and margin maximization.
- **Q6** Discuss the Kernel Trick. Compare Linear, Polynomial, and Gaussian RBF kernels.
- **Q7** What are the major advantages and limitations of SVM in high-dimensional spaces?

Unit 2 done. You are exam-ready.

Revise the Exam corner once more, then open Unit 3 on theycotes.com. Every AKTU subject, unit by unit, free. Notes by Chirag Jain, Ycotes.

