Machine Learning - List of questions

Learning Theory

Beginner Level

Intermediate Level

1. What is the VC dimension?
2. What is the VC dimension for an n-dimensional linear classifier?
3. What does the training set size depend on for a finite and infinite hypothesis set? Compare and contrast.
4. From the bias-variance trade-off, can you derive the bound-on training set size?
5. How is the VC dimension of a SVM bounded although it is projected to an infinite dimension?

Advanced Level

1. What is Union bound and Hoeffding's inequality?
2. State the uniform convergence theorem and derive it.
3. What is the sample complexity bound of the uniform convergence theorem?
4. What is the error bound of the uniform convergence theorem?
5. Considering that Empirical Risk Minimization is a NP-hard problem, how does logistic regression and SVM loss work?

1. Describe bias and variance with examples.

Bias refers to the error introduced by approximating a real-world problem, which may be complex, by a simplified model. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).

Variance refers to the error introduced by the model's sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data rather than the intended outputs (overfitting).

Examples:

- **High Bias, Low Variance:** A linear regression model on a dataset with a clear non-linear relationship. The model is too simple and does not capture the complexity of the data, resulting in high bias.
- **Low Bias, High Variance:** A polynomial regression model with a high degree on the same dataset. The model fits the training data very well but may perform poorly on new data due to overfitting, resulting in high variance.

2. What is Empirical Risk Minimization?

Empirical Risk Minimization (ERM) is a principle in statistical learning theory that aims to minimize the average loss on a given sample of data. The empirical risk is the average of the loss function over the training set, and ERM looks to find the model parameters that minimize this empirical risk.

Mathematically, if we have a loss function LLL and training data (x1,y1),…,(xn,yn)(x_1, y_1), \ldots, (x_n, y_n)(x1,y1),…,(xn,yn), ERM solves the following optimization problem:

min⁡θ1n∑i=1nL(yi, f(xi;θ))\min_{\theta} \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i; \theta))minθn1∑i=1nL(yi,f(xi;θ))

where f(xi;θ)f(x_i; \theta)f(xi;θ) is the prediction of the model with parameters θ\thetaθ.

3. Write the formula for training error and generalization error. Point out the differences.

Training Error: The error of the model on the training dataset. It is the average loss over the training set.

Training Error=1n∑i=1nL(yi,y^i)\text{Training Error} = \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{y}_i)Training Error=n1∑i=1nL(yi,y^i) where y^i\hat{y}_iy^i is the prediction for the iii-th training example.

Generalization Error: The error of the model on an independent test dataset that was not seen during training. It is the expected loss over the distribution of the data.

Generalization Error=E(x,y)∼P[L(y,f(x))]\text{Generalization Error} = \mathbb{E}_{(x,y) \sim P} [L(y, f(x))]Generalization Error=E(x,y)∼P[L(y,f(x))]

Differences:

- **Training Error:** Measured on the training data, may be lower because the model has been optimized for this specific data.
- **Generalization Error:** Measured on new, unseen data, provides an estimate of how well the model will perform on real-world data.

4. What is the bias-variance trade-off theorem?

The **bias-variance trade-off theorem** describes the trade-off between two sources of error that affect the performance of a machine learning model:

- **Bias:** Error due to overly simplistic assumptions in the learning algorithm.
- **Variance:** Error due to too much complexity in the learning algorithm.

The theorem states that as you increase the complexity of the model, the bias decreases and the variance increases, and vice versa. The goal is to find a balance between bias and variance to minimize the total error.

Total Error = Bias^2 + Variance + Irreducible Error

- **High Bias, Low Variance:** Underfitting, where the model is too simple.
- **Low Bias, High Variance:** Overfitting, where the model is too complex.

Finding the right balance is crucial for developing a model that generalizes well to new data.

5. What is the VC dimension?

The **VC dimension** (Vapnik-Chervonenkis dimension) is a measure of the capacity or complexity of a hypothesis space in statistical learning theory. It is defined as the maximum number of points that can be shattered (i.e., correctly classified in all possible ways) by the hypothesis space. A hypothesis space with a higher VC dimension can represent more complex functions.

What is the VC dimension for an n-dimensional linear classifier?

For an nnn-dimensional linear classifier, the VC dimension is n+1n + 1n+1. This means that a linear classifier in nnn dimensions can shatter any set of n+1n+1n+1 points in general position.

6. What does the training set size depend on for a finite and infinite hypothesis set? Compare and contrast.

Finite Hypothesis Set:

- The training set size required to achieve a certain level of accuracy and confidence depends on the size of the hypothesis set.
- The generalization error can be bounded using the union bound over all hypotheses.

Infinite Hypothesis Set:

- The training set size required is influenced by the complexity of the hypothesis space, often measured by the VC dimension.
- For infinite hypothesis sets, bounds on the generalization error are given in terms of the VC dimension rather than the number of hypotheses.

Comparison:

- For finite hypothesis sets, the dependency is directly on the number of hypotheses, whereas for infinite sets, it is on the complexity (VC dimension).
- Infinite hypothesis sets typically require more data to achieve the same level of generalization error due to their higher complexity.

7. From the bias-variance trade-off, can you derive the bound on training set size?

The bias-variance trade-off provides insight into the relationship between model complexity and error, but deriving a bound on the training set size involves more detailed theoretical analysis. The training set size mmm needed for a certain generalization error ϵ\epsilonϵ with confidence 1−δ1 - \delta1−δ can be approximated by:

m=O(VC dimensionϵlog⁡(1ϵ)+1ϵlog⁡(1δ))m = O\left(\frac{\text{VC dimension}}{\epsilon} \log\left(\frac{1}{\epsilon}\right) + \frac{1}{\epsilon} \log\left(\frac{1}{\delta}\right)\right)m=O(ϵVC dimensionlog(ϵ1)+ϵ1log(δ1))

This bound shows that as the VC dimension increases, the required training set size also increases to ensure good generalization.

8. How is the VC dimension of a SVM bounded although it is projected to an infinite dimension?

Support Vector Machines (SVMs) can operate in high-dimensional or even infinite-dimensional feature spaces using kernel functions. Despite this, the VC dimension of an SVM is bounded because it depends on the margin by which the data points can be separated rather than the dimensionality of the feature space.

The key factors that bound the VC dimension of an SVM are:

- The margin: Larger margins lead to a lower VC dimension.

- The number of support vectors: The VC dimension is influenced by the number of support vectors, which in turn is related to the margin.

Mathematically, for linearly separable data, the VC dimension ddd of an SVM can be bounded by:

d≤min(R2γ2, n)d \leq \min \left(\frac{R^2}{\gamma^2}, n \right)d≤min(γ2R2,n)

where RRR is the radius of the smallest sphere enclosing the data, γ\gammaγ is the margin, and nnn is the number of training samples.

This ensures that SVMs maintain good generalization properties even when projected into high or infinite-dimensional spaces using kernels.

9. What is Union Bound and Hoeffding's Inequality?

Union Bound: The union bound is a fundamental principle in probability theory that states the probability of at least one of a number of events occurring is less than or equal to the sum of the probabilities of the individual events. Mathematically, for events A1,A2,…,AnA_1, A_2, \ldots, A_nA1,A2,…,An:

P(⋃i=1nAi)≤∑i=1nP(Ai)\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathbb{P}(A_i)P(⋃i=1nAi)≤∑i=1nP(Ai)

Hoeffding's Inequality: Hoeffding's inequality provides an upper bound on the probability that the sum of bounded independent random variables deviates from its expected value by more than a certain amount. Specifically, if X1,X2,…,XnX_1, X_2, \ldots, X_nX1,X2,…,Xn are independent random variables bounded by the interval [a,b][a, b][a,b], then for any t>0t > 0t>0:

P(∣1n∑i=1nXi−E[1n∑i=1nXi]∣≥t)≤2exp(−2nt2(b−a)2)\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right]\right| \geq t \right) \leq 2 \exp\left(-\frac{2nt^2}{(b-a)^2}\right)P(n1∑i=1nXi−E[n1∑i=1nXi] ≥t)≤2exp(−(b−a)22nt2)

10. State the Uniform Convergence Theorem and Derive It

Uniform Convergence Theorem: The uniform convergence theorem states that for a hypothesis class H\mathcal{H}H with finite VC dimension ddd, the empirical risk converges uniformly to the true risk as the sample size increases. This means that with high probability, the difference between the empirical risk and the true risk is small for all hypotheses in H\mathcal{H}H.

Derivation: Given a hypothesis class H\mathcal{H}H with VC dimension ddd, let S={(x1,y1),…,(xn,yn)}S = \{(x_1, y_1), \ldots, (x_n, y_n)\}S={(x1,y1),…,(xn,yn)} be a training set of size nnn drawn i.i.d. from a distribution PPP. Let L(h)L(h)L(h) be the true risk and L^(h)\hat{L}(h)L^(h) be the empirical risk for hypothesis hhh.

By Hoeffding's inequality, for a single hypothesis hhh:

P(∣L(h)−L^(h)∣≥ϵ)≤2exp⁡(−2nϵ2)\mathbb{P}\left(\left| L(h) - \hat{L}(h) \right| \geq \epsilon \right) \leq 2 \exp\left(-2n\epsilon^2 \right)P(L(h)−L^(h)≥ϵ)≤2exp(−2nϵ2)

To apply the union bound to all hypotheses in H\mathcal{H}H:

P(∃h∈H:∣L(h)−L^(h)∣≥ϵ)≤∣H∣⋅2exp⁡(−2nϵ2)\mathbb{P}\left(\exists h \in \mathcal{H} : \left| L(h) - \hat{L}(h) \right| \geq \epsilon \right) \leq |\mathcal{H}| \cdot 2 \exp\left(-2n\epsilon^2 \right)P(∃h∈H:L(h)−L^(h)≥ϵ)≤∣H∣⋅2exp(−2nϵ2)

For a finite hypothesis class with VC dimension ddd, the number of hypotheses ∣H∣|\mathcal{H}|∣H∣ can be bounded by O(nd)O(n^d)O(nd). Hence:

P(∃h∈H:∣L(h)−L^(h)∣≥ϵ)≤2ndexp⁡(−2nϵ2)\mathbb{P}\left(\exists h \in \mathcal{H} : \left| L(h) - \hat{L}(h) \right| \geq \epsilon \right) \leq 2 n^d \exp\left(-2n\epsilon^2 \right)P(∃h∈H:L(h)−L^(h)≥ϵ)≤2ndexp(−2nϵ2)

Setting this probability to a small value δ\deltaδ:

2ndexp⁡(−2nϵ2)≤δ2 n^d \exp\left(-2n\epsilon^2 \right) \leq \delta2ndexp(−2nϵ2)≤δ

Solving for nnn:

n≥12ϵ2(dlog⁡n+log⁡2δ)n \geq \frac{1}{2\epsilon^2} \left(d \log n + \log \frac{2}{\delta} \right)n≥2ϵ21(dlogn+logδ2)

This gives the bound on the sample size nnn for uniform convergence.

11. What is the Sample Complexity Bound of the Uniform Convergence Theorem?

The sample complexity bound of the uniform convergence theorem is the number of samples nnn required to ensure that the empirical risk uniformly converges to the true risk with high probability 1−δ1 - \delta1−δ. Given a hypothesis class H\mathcal{H}H with VC dimension ddd and desired accuracy ϵ\epsilonϵ, the sample complexity nnn is:

n≥O(dlog⁡n+log⁡1δϵ2)n \geq O\left(\frac{d \log n + \log \frac{1}{\delta}}{\epsilon^2} \right)n≥O(ϵ2dlogn+logδ1)

12. What is the Error Bound of the Uniform Convergence Theorem?

The error bound of the uniform convergence theorem states that with high probability 1−δ1 - \delta1−δ, the difference between the true risk L(h)L(h)L(h) and the empirical risk L^(h)\hat{L}(h)L^(h) for all hypotheses h∈Hh \in \mathcal{H}h∈H is bounded by:

∣L(h)−L^(h)∣≤O(dlog⁡n+log⁡1δn)\left| L(h) - \hat{L}(h) \right| \leq O\left(\sqrt{\frac{d \log n + \log \frac{1}{\delta}}{n}} \right)L(h)−L^(h)≤O(ndlogn+logδ1)

13. Considering that Empirical Risk Minimization is a NP-hard problem, how does Logistic Regression and SVM Loss Work?

Logistic Regression:

- **Loss Function:** Logistic regression uses the logistic loss (or cross-entropy loss) to measure the error. The loss function for a single training example (xi,yi)(x_i, y_i)(xi,yi) is:

  L(θ;xi,yi)=−yilog⁡(hθ(xi))−(1−yi)log⁡(1−hθ(xi))L(\theta; x_i, y_i) = - y_i \log(h_\theta(x_i)) - (1 - y_i) \log(1 - h_\theta(x_i))L(θ;xi,yi)=−yilog(hθ(xi))−(1−yi)log(1−hθ(xi))

  where hθ(xi)h_\theta(x_i)hθ(xi) is the predicted probability that yi=1y_i = 1yi=1.

- **Optimization:** The logistic loss is convex, allowing the use of efficient optimization algorithms like gradient descent to find the parameters θ\thetaθ that minimize the empirical risk.

Support Vector Machines (SVM):

- **Loss Function:** SVMs use the hinge loss for binary classification. The loss function for a single training example (xi,yi)(x_i, y_i)(xi,yi) is:

  L(θ;xi,yi)=max⁡(0,1−yi(θ⋅xi))L(\theta; x_i, y_i) = \max(0, 1 - y_i (\theta \cdot x_i))L(θ;xi,yi)=max(0,1−yi(θ⋅xi))

- **Optimization:** The SVM optimization problem involves minimizing the hinge loss with a regularization term to control model complexity. This results in a convex optimization problem that can be solved efficiently using techniques like the Sequential Minimal Optimization (SMO) algorithm or gradient descent.

Both logistic regression and SVMs transform the NP-hard ERM problem into convex optimization problems, which can be solved efficiently with existing algorithms.