Week 00a Advanced Handout: ML Foundations - Mathematical Theory

Machine Learning for Smarter Innovation

1 Week 00a Advanced Handout: ML Foundations - Mathematical Theory

1.1 For Students With: Calculus, linear algebra, probability

1.2 Statistical Learning Theory

1.2.1 Empirical Risk Minimization

Goal: Minimize expected loss over data distribution

$$R(f) = \mathbb{E}_{(x,y)\sim P}[L(f(x), y)]$$

Empirical Risk (training error):

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i)$$

1.2.2 Bias-Variance Decomposition

For squared loss:

$$\mathbb{E}[(y - \hat{f}(x))^2] = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

- **Bias**: Error from wrong assumptions (underfitting)
- **Variance**: Error from sensitivity to training set (overfitting)
- **Tradeoff**: Simple models (high bias, low variance), Complex models (low bias, high variance)

1.2.3 VC Dimension and Generalization

VC Dimension (Vapnik-Chervonenkis): Measure of model capacity

Generalization Bound:

$$R(f) \leq \hat{R}(f) + \sqrt{\frac{d(\log(2n/d) + 1) - \log(\delta/4)}{n}}$$

Where: - d = VC dimension - n = training samples - δ = confidence level

Insight: Generalization improves with more data, degrades with model complexity

---

1.3 PAC Learning Framework

1.3.1 Probably Approximately Correct (PAC) Learning

A concept class is PAC-learnable if:

Given: - Accuracy $\epsilon > 0$ - Confidence $\delta > 0$ - Sample complexity $m(\epsilon, \delta)$

Algorithm outputs hypothesis h with:

$$P(R(h) \leq \epsilon) \geq 1 - \delta$$

1.3.2 Sample Complexity Bounds

For finite hypothesis space $|\mathcal{H}|$:

$$m \geq \frac{1}{\epsilon}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)$$

Implication: Need more samples for complex models

1.4 Optimization Theory

1.4.1 Gradient Descent Convergence

For convex L-Lipschitz functions with learning rate $\eta = \frac{1}{\sqrt{t}}$:

$$R(w_t) - R(w^*) \leq \frac{L\|w_0 - w^*\|}{\sqrt{t}}$$

Converges at rate $O(1/\sqrt{t})$

1.4.2 Stochastic Gradient Descent (SGD)

Update rule:

$$w_{t+1} = w_t - \eta_t \nabla L(w_t; x_i, y_i)$$

Advantages: - Faster per-iteration (single sample) - Escapes local minima (noise helps) - Online learning compatible

Convergence: $O(1/\sqrt{t})$ with proper learning rate schedule

---

1.5 Regularization Theory

1.5.1 Ridge Regression (L2)

$$\min_w \|Xw - y\|^2 + \lambda\|w\|^2$$

Closed Form:

$$w^* = (X^T X + \lambda I)^{-1} X^T y$$

Bayesian Interpretation: Gaussian prior on weights

1.5.2 Lasso (L1)

$$\min_w \|Xw - y\|^2 + \lambda \|w\|_1$$

Properties: - Sparse solutions (many $w_i = 0$) - Feature selection - No closed form (use proximal gradient)

1.5.3 Elastic Net

$$\min_w \|Xw - y\|^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|^2$$

Combines L1 sparsity with L2 stability

1.6 Kernel Methods

1.6.1 Kernel Trick

Implicit high-dimensional mapping via kernel function:

$$K(x, x') = \langle \phi(x), \phi(x') \rangle$$

Common Kernels: 1. **Linear**: $K(x, x') = x^T x'$ 2. **Polynomial**: $K(x, x') = (x^T x' + c)^d$ 3. **RBF**: $K(x, x') = \exp(-\gamma \|x - x'\|^2)$

1.6.2 Representer Theorem

Optimal solution lives in span of training data:

$$f^*(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$$

Implication: Can solve in dual space (useful when $d \gg n$)

1.7 Information Theory in ML

1.7.1 Cross-Entropy Loss

$$H(p, q) = -\sum_i p_i \log q_i$$

Classification: Minimizing cross-entropy = maximizing likelihood

1.7.2 KL Divergence

$$D_{KL}(P \| Q) = \sum_i p_i \log \frac{p_i}{q_i}$$

Properties: - Non-negative - Zero iff $P = Q$ - Asymmetric

Use: Measure distribution mismatch (VAEs, RL)

1.7.3 Mutual Information

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Application: Feature selection (maximize $I(X_i;Y)$)

1.8 Concentration Inequalities

1.8.1 Hoeffding's Inequality

$$P(|\hat{\mu} - \mu| \geq \epsilon) \leq 2\exp(-2n\epsilon^2)$$

Application: Confidence intervals for empirical mean

1.8.2 McDiarmid's Inequality

For bounded differences c_i:

$$P(|f - \mathbb{E}[f]| \geq \epsilon) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum c_i^2}\right)$$

Application: Generalization bounds for more complex functions

---

1.9 Advanced Topics

1.9.1 1. Online Learning

Regret Bound:

$$\text{Regret}_T = \sum_{t=1}^{T} L(w_t, x_t, y_t) - \min_w \sum_{t=1}^{T} L(w, x_t, y_t)$$

Goal: Sublinear regret $o(T)$

1.9.2 2. Multi-Armed Bandits

Explore-Exploit Tradeoff: Upper Confidence Bound (UCB)

$$a_t =_a \left(\hat{\mu}_a + \sqrt{\frac{2\log t}{n_a}}\right)$$

1.9.3 3. Boosting Theory

AdaBoost minimizes exponential loss:

$$L(\alpha, w) = \sum_i \exp(-y_i f(x_i))$$

Margin Theory: Boosting increases minimum margin

1.10 Proofs

1.10.1 Proof: Ridge Regression Solution

Minimize:

$$L(w) = \|Xw - y\|^2 + \lambda\|w\|^2$$

Take gradient:

$$\nabla_w L = 2X^T(Xw - y) + 2\lambda w = 0$$

Solve for w:

$$X^T Xw + \lambda w = X^T y$$
$$(X^T X + \lambda I)w = X^T y$$
$$w^* = (X^T X + \lambda I)^{-1} X^T y$$

Note: λI ensures invertibility even when $X^T X$ singular

1.11 Practice Problems

1. **Derive** logistic regression gradient
2. **Prove** bias-variance decomposition for squared loss
3. **Show** SVMs solve dual problem using KKT conditions
4. **Compute** VC dimension of linear classifiers in $\mathbb{R}^d$
5. **Analyze** convergence rate of batch vs stochastic gradient descent

---

1.12 References

- Shalev-Shwartz & Ben-David: Understanding Machine Learning
- Hastie, Tibshirani, Friedman: Elements of Statistical Learning
- Vapnik: Statistical Learning Theory
- Bishop: Pattern Recognition and Machine Learning