Foundational Data Science Summary

Mathematical Notes

October 27, 2025

Contents

1 Statistical Learning Theory

1.1 Learning Problem

Definition 1.1. A **learning problem** consists of:

- Input space $\mathcal{X}$
- Output space $\mathcal{Y}$
- Training data $S = \{(x_i, y_i)\}_{i=1}^n$ where $(x_i, y_i) \sim P$
- Hypothesis class $\mathcal{H}$
- Loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

1.2 Empirical Risk Minimization

Definition 1.2. The **empirical risk** is:

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

Definition 1.3. The **true risk** is:

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

1.3 Generalization Error

Definition 1.4. The **generalization error** is:

$$\epsilon(h) = R(h) - \hat{R}(h)$$

1.4 VC Dimension

Definition 1.5. The **VC dimension** of a hypothesis class $\mathcal{H}$ is the largest number d such that $\mathcal{H}$ can shatter any set of d points.

Theorem 1.1 (VC Bound)**.** With probability at least $1 - \delta$:

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

where d is the VC dimension of $\mathcal{H}$.

2 Exponential Families

2.1 Definition

Definition 2.1. A probability distribution belongs to an **exponential family** if its density can be written as:

$$p(x|\theta) = h(x) \exp(\eta(\theta)^T T(x) - A(\theta))$$

where:

- $h(x)$ is the base measure
- $\eta(\theta)$ is the natural parameter
- $T(x)$ is the sufficient statistic
- $A(\theta)$ is the log-partition function

2.2 Canonical Form

Definition 2.2. The **canonical form** of an exponential family is:

$$p(x|\eta) = h(x)\exp(\eta^T T(x) - A(\eta))$$

2.3 Properties

Theorem 2.1. For exponential families:

- $\mathbb{E}[T(X)] = \nabla A(\eta)$
- $\mathrm{Cov}[T(X)] = \nabla^2 A(\eta)$
- $A(\eta)$ is convex

2.4 Examples

- **Bernoulli**: $p(x|p) = p^x(1-p)^{1-x}$
- **Poisson**: $p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$
- **Gaussian**: $p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$
- **Multinomial**: $p(x|p) = \frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k}$

3 Maximum Likelihood Estimation

3.1 Likelihood Function

Definition 3.1. The **likelihood function** is:

$$L(\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

Definition 3.2. The **log-likelihood function** is:

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log p(x_i|\theta)$$

3.2 MLE for Exponential Families

Theorem 3.1. For exponential families, the MLE satisfies:

$$\nabla A(\hat{\eta}) = \frac{1}{n}\sum_{i=1}^{n} T(x_i)$$

3.3 Fisher Information

Definition 3.3. The **Fisher information matrix** is:

$$I(\theta) = \mathbb{E}\left[-\frac{\partial^2 \log p(X|\theta)}{\partial\theta\partial\theta^T}\right]$$

Theorem 3.2 (Cramér-Rao Lower Bound)**.** For any unbiased estimator $\hat{\theta}$:

$$\text{Var}(\hat{\theta}) \geq I(\theta)^{-1}$$

4 Bayesian Inference

4.1 Bayes' Theorem

Theorem 4.1 (Bayes' Theorem)**.**

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}$$

4.2 Conjugate Priors

Definition 4.1. A prior $p(\theta)$ is **conjugate** to a likelihood $p(x|\theta)$ if the posterior $p(\theta|x)$ belongs to the same family as the prior.

4.3 Examples of Conjugate Pairs

- **Beta-Bernoulli**: Beta prior for Bernoulli likelihood
- **Dirichlet-Multinomial**: Dirichlet prior for multinomial likelihood
- **Normal-Normal**: Normal prior for normal likelihood
- **Gamma-Poisson**: Gamma prior for Poisson likelihood

4.4 Posterior Predictive Distribution

Definition 4.2. The **posterior predictive distribution** is:

$$p(x_{new}|x) = \int p(x_{new}|\theta)p(\theta|x)d\theta$$

5 Multi-Armed Bandits

5.1 Problem Setup

Definition 5.1. A **multi-armed bandit** problem consists of:

- K arms (actions)
- At each time t, choose arm $a_t \in \{1, \ldots, K\}$
- Receive reward $r_t \sim P_{a_t}$
- Goal: maximize $\sum_{t=1}^{T} r_t$

5.2 Regret

Definition 5.2. The **cumulative regret** is:

$$R_T = \sum_{t=1}^{T} (\mu^* - \mu_{a_t})$$

where $\mu^* = \max_i \mu_i$ and $\mu_i = \mathbb{E}[r|a = i]$.

5.3 Upper Confidence Bound (UCB)

Definition 5.3. The **UCB1 algorithm** selects arm:

$$a_t = \arg\max_i \left(\hat{\mu}_i + c\sqrt{\frac{\log t}{n_i}} \right)$$

where $\hat{\mu}_i$ is the empirical mean of arm i, n_i is the number of times arm i has been pulled, and c is a constant.

Theorem 5.1 (UCB Regret Bound)**.** The UCB1 algorithm achieves:

$$R_T \leq 8 \sum_{i:\mu_i<\mu^*} \frac{\log T}{\Delta_i} + \left(1 + \frac{\pi^2}{3}\right) \sum_{i=1}^{K} \Delta_i$$

where $\Delta_i = \mu^* - \mu_i$.

5.4 Thompson Sampling

Definition 5.4. Thompson sampling selects arm a_t according to:

$$a_t \sim \text{argmax}_i \theta_i^{(t)}$$

where $\theta_i^{(t)} \sim p(\theta_i|\mathcal{H}_{i,t})$ is sampled from the posterior distribution of arm i.

5.5 Contextual Bandits

Definition 5.5. In **contextual bandits**, at each time t:

- Observe context $x_t \in \mathbb{R}^d$
- Choose arm $a_t \in \{1, \ldots, K\}$
- Receive reward $r_t = f_{a_t}(x_t) + \epsilon_t$

5.6 Linear Contextual Bandits

Definition 5.6. In **linear contextual bandits**, the expected reward is:

$$\mathbb{E}[r_t|a_t, x_t] = \theta_{a_t}^T x_t$$

where $\theta_{a_t} \in \mathbb{R}^d$ is the parameter vector for arm a_t.

6 Online Learning

6.1 Online Convex Optimization

Definition 6.1. In **online convex optimization**:

- At time t, choose $w_t \in \mathcal{W}$
- Observe convex loss function $f_t : \mathcal{W} \to \mathbb{R}$
- Suffer loss $f_t(w_t)$
- Goal: minimize $\sum_{t=1}^{T} f_t(w_t)$

6.2 Regret in Online Learning

Definition 6.2. The **regret** is:

$$R_T = \sum_{t=1}^{T} f_t(w_t) - \min_{w \in \mathcal{W}} \sum_{t=1}^{T} f_t(w)$$

6.3 Gradient Descent

Theorem 6.1 (Online Gradient Descent)**.** For convex functions with bounded gradients, online gradient descent with step size $\eta_t = \frac{1}{\sqrt{t}}$ achieves:

$$R_T \leq O(\sqrt{T})$$

6.4 Follow the Regularized Leader (FTRL)

Definition 6.3. FTRL chooses:

$$w_{t+1} = \arg\min_{w \in \mathcal{W}} \left(\sum_{s=1}^{t} f_s(w) + R(w) \right)$$

where $R(w)$ is a regularization term.

7 Information Theory

7.1 Entropy

Definition 7.1. The **Shannon entropy** of a discrete random variable X is:

$$H(X) = -\sum_{x} p(x) \log p(x)$$

Definition 7.2. The **differential entropy** of a continuous random variable X is:

$$h(X) = -\int p(x) \log p(x) dx$$

7.2 Mutual Information

Definition 7.3. The **mutual information** between random variables X and Y is:

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

7.3 Kullback-Leibler Divergence

Definition 7.4. The **KL divergence** between distributions P and Q is:

$$D_{KL}(P||Q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

7.4 Cross-Entropy

Definition 7.5. The **cross-entropy** between distributions P and Q is:

$$H(P,Q) = -\sum_x p(x) \log q(x) = H(P) + D_{KL}(P||Q)$$

8 Dimension Reduction

8.1 Principal Component Analysis (PCA)

Definition 8.1. PCA finds the linear transformation that maximizes the variance of the projected data:

$$\max_{\mathbf{w}} \mathbf{w}^T \mathbf{S} \mathbf{w} \quad \text{subject to} \quad \|\mathbf{w}\| = 1$$

where $\mathbf{S}$ is the covariance matrix.

8.2 Singular Value Decomposition

Theorem 8.1. Any matrix $A \in \mathbb{R}^{m \times n}$ can be decomposed as:

$$A = U \Sigma V^T$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal, and $\Sigma \in \mathbb{R}^{m \times n}$ is diagonal.

8.3 Linear Discriminant Analysis (LDA)

Definition 8.2. LDA finds the linear transformation that maximizes the ratio of between-class to within-class variance:

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

where $\mathbf{S}_B$ is the between-class scatter matrix and $\mathbf{S}_W$ is the within-class scatter matrix.

9 Clustering

9.1 K-Means

Definition 9.1. K-means minimizes:

$$J = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2$$

where $w_{ik} = 1$ if $\mathbf{x}_i$ belongs to cluster k, 0 otherwise.

9.2 Gaussian Mixture Models

Definition 9.2. A **Gaussian mixture model** has density:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where π_k are mixing weights and $\sum_{k=1}^{K} \pi_k = 1$.

9.3 Expectation-Maximization (EM)

Theorem 9.1 (EM Algorithm)**.** For mixture models, EM alternates between:

- **E-step**: Compute $q(z_{ik}) = p(z_i = k|\mathbf{x}_i, \theta^{(t)})$
- **M-step**: Update $\theta^{(t+1)} = \arg\max_\theta \sum_{i,k} q(z_{ik}) \log p(\mathbf{x}_i, z_i = k|\theta)$

10 Graphical Models

10.1 Bayesian Networks

Definition 10.1. A **Bayesian network** is a directed acyclic graph where each node represents a random variable and edges represent conditional dependencies.

Theorem 10.1 (Factorization)**.** For a Bayesian network, the joint probability factors as:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\text{pa}(x_i))$$

where $\text{pa}(x_i)$ are the parents of x_i.

10.2 Markov Random Fields

Definition 10.2. A **Markov random field** is an undirected graph where each node represents a random variable and edges represent dependencies.

10.3 Inference

- **Variable elimination**
- **Belief propagation**
- **Markov chain Monte Carlo (MCMC)**
- **Variational inference**

11 Time Series Analysis

11.1 Stationarity

Definition 11.1. A time series $\{X_t\}$ is **weakly stationary** if:

- $\mathbb{E}[X_t] = \mu$ (constant mean)
- $\text{Cov}(X_t, X_{t+h}) = \gamma(h)$ (covariance depends only on lag h)

11.2 ARIMA Models

Definition 11.2. An **ARIMA**(p,d,q) model is:

$$\phi(B)(1-B)^d X_t = \theta(B)\epsilon_t$$

where B is the backshift operator, $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$, and $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$.

11.3 Kalman Filter

Definition 11.3. The **Kalman filter** provides optimal estimates for linear state-space models:

$$\mathbf{x}_t = F\mathbf{x}_{t-1} + G\mathbf{u}_t + \mathbf{w}_t \tag{1}$$
$$\mathbf{y}_t = H\mathbf{x}_t + \mathbf{v}_t \tag{2}$$

where $\mathbf{w}_t \sim \mathcal{N}(0, Q)$ and $\mathbf{v}_t \sim \mathcal{N}(0, R)$.

12 Causal Inference

12.1 Rubin Causal Model

Definition 12.1. The **Rubin causal model** defines:

- Potential outcomes: $Y_i(1)$ and $Y_i(0)$
- Treatment indicator: $T_i \in \{0, 1\}$
- Observed outcome: $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$

12.2 Average Treatment Effect

Definition 12.2. The **average treatment effect** is:

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)]$$

12.3 Instrumental Variables

Definition 12.3. An **instrumental variable** Z satisfies:

- Relevance: $\text{Cov}(Z, T) \neq 0$
- Exogeneity: $\text{Cov}(Z, U) = 0$ where U contains unobserved confounders

13 Applications

13.1 Recommendation Systems

- Collaborative filtering
- Content-based filtering
- Matrix factorization
- Deep learning approaches

13.2 A/B Testing

- Statistical significance testing
- Power analysis
- Multiple testing corrections
- Sequential testing

13.3 Anomaly Detection

- Statistical methods
- Machine learning approaches
- Time series methods
- Graph-based methods

14 Important Algorithms

14.1 Optimization

- Gradient descent
- Stochastic gradient descent
- Adam optimizer
- Newton's method

14.2 Sampling

- Metropolis-Hastings
- Gibbs sampling
- Importance sampling
- Rejection sampling

14.3 Regularization

- Ridge regression
- Lasso regression
- Elastic net
- Dropout

15 Key Theorems

15.1 Central Limit Theorem

Theorem 15.1 (Central Limit Theorem)**.** If $X_1, X_2, \ldots$ are i.i.d. with mean μ and variance σ^2, then:

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$$

15.2 Law of Large Numbers

Theorem 15.2 (Strong Law of Large Numbers)**.** If $X_1, X_2, \ldots$ are i.i.d. with $\mathbb{E}[X_i] = \mu < \infty$, then:

$$\bar{X}_n \xrightarrow{a.s.} \mu$$

15.3 Universal Approximation Theorem

Theorem 15.3. A feedforward neural network with a single hidden layer can approximate any continuous function on a compact set, given sufficient hidden units.