Lecture 1: Introduction Probability Theory

Terminologies

- **Instance**: measurements about individual entities/objects (no label)
- **Attributes**: component of the instances
- **Label**: an outcome that is categorical, numerical, etc.
- **Examples**: instance coupled with label
- **Models**: discovered relationship between attributes and/or label

Supervised v.s. Unsupervised

- **Supervised**:
  - Labelled data
  - Predict labels on new instances
- **Unsupervised**:
  - Unlabelled data
  - Cluster related instances; project to fewer dimensions; understand attribute relationships

Evaluation

1. Pick an evaluation metric comparing label v.s. prediction
   - E.g. Accuracy, Contingency table, Precision-Recall, ROC curves
2. Procure an independent, labelled test set
3. "Average" the evaluation metric over the test set (When data poor, use cross-validation)

Probability相关的部分就不写了

Lecture 2: Statistical Schools of Thoughts

Frequentist statistics

- Unknown params are treated as having fixed but unknown values
- Parameter estimation:
  - Classes of models $\{p_\theta(x) : \theta \in \Theta\}$ indexed by parameters Θ
  - Point estimate $\hat{\theta}(x_1, ..., x_n)$: a function (or statistic) of data (samples)
    - A single value as an estimate of a population parameter
- If $\hat{\theta}$ is an estimator for θ
  - Bias: $Bias_\theta(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta$
  - Variance: $Var_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}])^2]$
- Asymptotic properties:
  - Consistency: $\hat{\theta} \to \theta$ (converges in probability) as $n \to \infty$
  - Efficiency: asymptotic variance is as small as possible (reach Cramer-Rao lower bound)
- Maximum-Likelihood Estimation (MLE)
  - General principle for designing estimators

- Involves optimisation
- $\hat{\theta} \in \underset{\theta \in \Theta}{\arg\max} \prod_{i=1}^{n} p_\theta(x_i)$
- MLE estimators are consistent (but usually biased)
- "Algorithm":
    1. Given data $X_1, ..., X_n$
    2. Likelihood: $L(\theta) = \prod_{i=1}^{n} p_\theta(X_i)$
    3. Optimise to find best params
        - Take partial derivatives of log likelihood: $l'(\theta)$
        - Solve $l'(\theta) = 0$
        - If fail, use iterative gradient method (e.g. fisher scoring)

Decision Theory

- Decision rule: $\delta(x) \in A$ (action space)
    - E.g. point estimate, out-of-sample prediction
- Loss function $l(a, \theta)$: economic cost, error metric
    - E.g. square loss $(\hat{\theta} - \theta)^2$, 0-1 loss $I(y \neq \hat{y})$

Risk & Empirical Risk Minimization (ERM)

- In decision theory, really care about **expected loss**
- **Risk** : $R_\theta[\delta] = E_{X \sim \theta}[l(\delta(X), \theta)]$
    - Risk = Expected Loss
    - aka. Generalization error
- **Goal**: Choose δ (decision) to minimise $R_\theta[\delta]$
    - Can't calculate risk directly
    - Don't know the real distribution the samples comes from, therefore don't now $E(X)$
- **ERM**
    - Use training set X to approximate p_θ (Empirical)
    - Minimise empirical risk $\hat{R}_\theta[\delta] = \frac{1}{n} \sum_{i=1}^{n} l(\delta(X_i), \theta)$

Mean Squared Error (of parameter estimator)

- Bias-variance decomposition of **square-loss risk**
- $E_\theta[(\theta - \hat{\theta})^2] = [Bias(\hat{\theta})]^2 + Var_\theta(\hat{\theta})$

Bayesian Statistics

- Unknown params have associated distributions reflecting prior **belief**
- Prior distribution $P(\theta)$
    - Params are modeled like r.v.'s
    - Data likelihood $P_\theta(X)$ written as conditional $P(X|\theta)$
- Rather than point estimate $\hat{\theta}$
    - Bayesians update prior belief $P(\theta)$ with observed data to the posterior distribution: $P(\theta|X)$
- Bayesian probabilistic inference
    1. Start with prior $P(\theta)$ and likelihood $P(X|\theta)$
    2. Observe data $X = x$

3. Update prior to posterior $P(\theta|X=x)$

- Primary tools to obtain the posterior
  - Bayes Rule: reverse order of conditioning
    - $P(\theta|X=x)=\frac{P(X=x|\theta)P(\theta)}{P(X=x)}$
  - Marginalization: eliminates unwanted variables
    - $P(X=x)=\sum_t P(X=x,\theta=t)$
- Bayesian point estimation common approaches
  - Posterior mean
    - $E_{\theta|X}[\theta]=\int \theta P(\theta|X)d\theta$
  - Posterior mode (MAP)
    - $\arg\max_\theta P(\theta|X)$
- MLE in Bayesian context
  - MLE = MAP if using uniform prior $P(\theta)$
  - (No prior belief about θ)

Categories of Probabilistic Models

- Parametric v.s. Non-Parametric
  1. Parametric
     - Determined by fixed, finite number of parameters
     - Limited flexibility
     - Efficient statistically and computationally
  2. Non-Parametric
     - Number of parameters grows with data, potentially infinite
     - More flexible
     - Less efficient
- Generative v.s. Discriminative
  1. Generative
     - Model full joint $P(X,Y)$
     - E.g. Naive Bayes
  2. Discriminative
     - Model conditional $P(Y|X)$ only (directly)
     - E.g. Linear Regression