Lecture 3: Statistical Decision Theory (Part II)

Hao Helen Zhang
Outline of This Note

- Part I: Statistics Decision Theory (Classical Statistical Perspective - “Estimation”)
  - loss and risk
  - MSE and bias-variance tradeoff
  - Bayes risk and minimax risk
- Part II: Learning Theory for Supervised Learning (Machine Learning Perspective - “Prediction”)
  - optimal learner
  - empirical risk minimization
  - restricted estimators
Supervised Learning

Any supervised learning problem has three components:

- Input vector $\mathbf{X} \in \mathbb{R}^d$.
- Output $Y$, either discrete or continuous valued.
- Probability framework

$$(\mathbf{X}, Y) \sim P(\mathbf{X}, Y).$$

The goal is to estimate the relationship between $\mathbf{X}$ and $Y$, described by $f(\mathbf{X})$, for future prediction and decision-making.

- For regression, $f : \mathbb{R}^d \rightarrow \mathbb{R}$
- For $K$-class classification, $f : \mathbb{R}^d \rightarrow \{1, \cdots, K\}$
Learning Loss Function

Similar to the learning theory, we use a *learning loss* function $L$ to measure the discrepancy $Y$ and $f(X)$ to penalize the errors for predicting $Y$.

Examples:

- squared error loss (used in mean regression)
  
  \[ L(Y, f(X)) = (Y - f(X))^2 \]

- absolute error loss (used in median regression)
  
  \[ L(Y, f(X)) = |Y - f(X)| \]

- 0-1 loss function (used in binary classification)
  
  \[ L(Y, f(X)) = I(Y \neq f(X)) \]
Risk, Expected Prediction Error (EPE)

The risk of $f$ is

$$R(f) = E_{X,Y} L(Y, f(X)) = \int L(y, f(x)) dP(x, y)$$

In machine learning, $R(f)$ is called Expected Prediction Error (EPE).

- For squared error loss,

  $$R(f) = EPE(f) = E_{X,Y} [Y - f(X)]^2$$

- For 0-1 loss

  $$R(f) = EPE(f) = E_{X,Y} I(Y \neq f(X)) = Pr(Y \neq f(X))$$

which is the probability of making a mistake.
We seek $f$ which minimizes the average (long-term) loss over the entire population.

**Optimal Learner:** The minimizer of $R(f)$ is

$$f^* = \arg \min_f R(f).$$

- The solution $f^*$ depends on the loss $L$.
- $f^*$ is not achievable without knowing $P(x, y)$ or the entire population

**Bayes Risk:** The optimal risk, or risk of the optimal learner

$$R(f^*) = E_{x, y}(Y - f^*(X))^2.$$
How to Find The Optimal Learner

Note that

\[ R(f) = E_{X,Y} L(Y, f(X)) = \int L(y, f(x)) dP(x, y) \]

\[ = E_X \{ E_{Y|X} L(Y, f(X)) \} \]

One can do the following,

- for each fixed \( x \), find \( f^*(x) \) by solving

\[ f^*(x) = \arg \min_c E_{Y|X=x} L(Y, c) \]
Example 1: Regression with Squared Error Loss

For regression, consider $L(Y, f(X)) = (Y - f(X))^2$. The risk is

$$R(f) = E_{X, Y}(Y - f(X))^2 = E_X E_{Y|X}((Y - f(X))^2|X).$$

- For each fixed $x$, the best learner solves the problem

$$\min_c E_{Y|X}[(Y - c)^2|X = x].$$

The solution is $E(Y|X = x)$, so

$$f^*(x) = E(Y|X = x) = \int y dP(y|x).$$

- So the optimal learner is $f^*(X) = E(Y|X)$. 

Hao Helen Zhang
Lecture 3: Statistical Decision Theory (Part II) 8 / 27
For any learner $f$, its EPE under the squared error loss is decomposed as

$$EPE(f) = E_{X,Y}[Y - f(X)]^2 = E_X E_{Y|X}([Y - f(X)]^2|X)$$

$$= E_X E_{Y|X}([Y - E(Y|X) + E(Y|X) - f(X)]^2|X)$$

$$= E_X E_{Y|X}([Y - E(Y|X)]^2|X) + E_X [f(X) - E(Y|X)]^2$$

$$= E_X Y [Y - f^*(X)]^2 + E_X [f(X) - f^*(X)]^2$$

$$= EPE(f^*) + MSE$$

$$= Bayes Risk + MSE$$

The Bayes risk (gold standard) gives a lower bound for any risk.
Example 2: Classification with 0-1 Loss

In binary classification, assume $Y \in \{0, 1\}$. The learning rule $f : R^d \rightarrow \{0, 1\}$.

Using 0-1 loss

$$L(Y, f(X)) = I(Y \neq f(X)),$$

the expected prediction error is

$$EPE(f) = EI(Y \neq f(X))$$

$$= \Pr(Y \neq f(X))$$

$$= E_X [P(Y = 1|X)I(f(X) = 0) + P(Y = 0|X)I(f(X) = 1)]$$
Bayes Classification Rule for Binary Problems

For each fixed $x$, the minimizer of $EPE(f)$ is given by

$$f^*(x) = \begin{cases} 
1 & \text{if } P(Y = 1|X = x) > P(Y = 0|X = x) \\
0 & \text{if } P(Y = 1|X = x) < P(Y = 0|X = x) 
\end{cases}$$

It is called the Bayes classifier, denoted by $\phi_B$.

The Bayes rule can be written as

$$\phi_B(x) = f^*(x) = I[P(Y = 1|X = x) > 0.5].$$

- To obtain the Bayes rule, we need to know $P(Y = 1|X = x)$, or whether $P(Y = 1|X = x = x)$ is larger than 0.5 or not.
Probability Framework

Assume the training set

$$(X_1, Y_1), \ldots, (X_n, Y_n) \sim \text{i.i.d } P(X, Y).$$

We aim to find the best function (optimal solution) from the model class $\mathcal{F}$ for decision making.

- However, the optimal learner $f^*$ is not attainable without knowledge of $P(x, y)$.
- The training set $(x_1, y_1), \ldots, (x_n, y_n)$ generated from $P(x, y)$ carry information about $P(x, y)$.
- We approximate the integral over $P(x, y)$ by the average over the training samples.
Empirical Risk Minimization

**Key Idea:** We cannot compute \( R(f) = E_{X,Y} L(Y, f(X)) \), but can compute its *empirical* version using the data!

*Empirical Risk:* also known as the *training error*

\[
R_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)).
\]

- A reasonable estimator \( f \) should be close to \( f^* \), or converge to \( f^* \) when more data are collected.

Using law of large numbers, \( R_{\text{emp}}(f) \to E_{\text{PE}}(f) \) as \( n \to \infty \).
Learning Based on ERM

We construct the estimator of $f^*$ by

$$
\hat{f} = \arg \min_f R_{\text{emp}}(f).
$$

This principle is called **Empirical Risk Minimization**. (ERM)

For regression,

- the empirical risk is known as **Residual Sum of Squares (RSS)**

$$
R_{\text{emp}}(f) = \text{RSS}(f) = \sum_{i=1}^{n} [y_i - f(x_i)]^2.
$$

- ERM gives the **least mean squares**:

$$
\hat{f}_{\text{ols}} = \arg \min_f \text{RSS}(f) = \arg \min_f \sum_{i=1}^{n} [y_i - f(x_i)]^2.
$$
Issues with ERM

Minimizing $R_{\text{emp}}(f)$ over all functions is not proper.

- Any function passing through all $(x_i, y_i)$ has $\text{RSS} = 0$.

Example:

$$f_1(x) = \begin{cases} y_i & \text{if } x = x_i \text{ for some } i = 1, \ldots, n \\ 1 & \text{otherwise.} \end{cases}$$

It is easy to check that $\text{RSS}(f_1) = 0$.

- However, it does not amount to any form of learning. For the future points not equal to $x_i$, it will always predict $y = 1$.

This phenomenon is called “overfitting”.

- Overfitting is undesired, since a small $R_{\text{emp}}(f)$ does not imply a small $R(f) = \int L(y, f(x))dP(x, y)$. 
Class of Restricted Estimators

We need to restrict the estimation process within a set of functions $\mathcal{F}$, to control \textit{complexity} of functions and hence avoid over-fitting.

- Choose a simple model class $\mathcal{F}$.

$$\hat{f}_{ols} = \arg \min_{f \in \mathcal{F}} \text{RSS}(f) = \arg \min_{f \in \mathcal{F}} \sum_{i=1}^{n} [y_i - f(x_i)]^2.$$ 

- Choose a large model class $\mathcal{F}$, but use a roughness penalty to control model complexity: Find $f \in \mathcal{F}$ to minimize

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \text{Penalized RSS} \equiv \text{RSS}(f) + \lambda J(f).$$

The solution is called \textbf{regularized empirical risk minimizer}.

How should we choose the model $\mathcal{F}$? Simple one or complex one?
Parametric Learners

Some model classes can be parametrized as

$$\mathcal{F} = \{ f(\mathbf{X}; \alpha), \alpha \in \Lambda \},$$

where $\alpha$ is the index parameter for the class.

**Parametric Model Class**: The form of $f$ is known except for finitely many unknown parameters

- **Linear Models**: linear regression, linear SVM, LDA
  
  $$f(\mathbf{X}; \beta) = \beta_0 + \mathbf{X}' \beta_1, \quad \beta_1 \in \mathbb{R}^d,$$
  
  $$f(\mathbf{X}; \beta) = \beta_0 + \sin(X_1)\beta_1 + \beta_2 X_2^2, \quad X \in \mathbb{R}^2$$

- **Nonlinear models**
  
  $$f(\mathbf{x}; \beta) = \beta_1 x^{\beta_2}$$
  
  $$f(\mathbf{X}; \beta) = \beta_1 + \beta_2 \exp^{\beta_3 \mathbf{X}^\beta_4}$$
Advantages of Linear Models

Why are linear models in wide use?

- If linear assumptions are correct, it is more efficient than nonparametric fits.
- Convenient and easy to fit
- Easy to interpret
- Providing the first-order Taylor approximation to $f(X)$

Example: In the linear regression, $\mathcal{F} = \{ \text{all linear functions of } x \}$.

The ERM is the same as solving the **ordinary least squares**, 

$$
(\hat{\beta}_0^{ols}, \hat{\beta}^{ols}) = \arg \min_{(\beta_0, \beta)} \sum_{i=1}^{n} [y_i - \beta_0 - x'_i \beta]^2.
$$
We never know whether linear assumptions are proper or not.

- Not flexible in practice, having all orders of derivatives and the same function form everywhere
- Individual observations have large influences on remote parts of the curve
- The polynomial degree can not be controlled continuously
More Flexible Learners

Nonparametric Model Class $\mathcal{F}$:
the function form $f$ is unspecified, and the parameter set could be infinitely dimensional.

- $\mathcal{F} = \{\text{all continuous functions of } x\}$.
- $\mathcal{F}$ is the second-order Sobolev space over $[0, 1]$.
- $\mathcal{F}$ is the space spanned by a set of Basis functions (dictionary functions)
  $$\mathcal{F} = \{ f : f_\theta(x) = \sum_{m=1}^{\infty} \theta_m h_m(x) \}.$$  

- Kernel methods and local regression:
  $\mathcal{F}$ contains all piece-wise constant functions,
or, $\mathcal{F}$ contains all piece-wise linear functions.
Example of Nonparametric Methods

For regression problems,
- splines, wavelets, kernel estimators, locally weighted polynomial regression, GAM
- projection pursuit, regression trees, k-nearest neighbors

For classification problems,
- kernel support vector machines, k-nearest neighbors
- classification trees ...

For density estimation,
- histogram, kernel density estimation
Semiparametric Models: a compromise between linear models and nonparametric models

- partially linear models
Nonparametric Models

Motivation: The underlying regression function is so complicated that no reasonable parametric models would be adequate.

Advantages:
- **NOT** assume any specific form.
- Assume less, thus discover more
- infinite parameters: not no parameters
- Local more relying on the neighbors
- Outliers and influential points do not affect the results

Let the data speak for themselves and show us the appropriate functional form!

Price we pay: more computational effort, lower efficiency and slower convergence rate (if linear models happen to be reasonable)
Nonparametric vs Parametric Models

They are not mutually exclusive competitors

- A nonparametric regression estimate may suggest a simple parametric model
- In practice, parametric models are preferred for simplicity
- Linear regression line is *infinitely smooth* fit

<table>
<thead>
<tr>
<th>Method</th>
<th># Parameters</th>
<th>Estimate</th>
<th>Outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>parametric</td>
<td>finite</td>
<td>global</td>
<td>sensitive</td>
</tr>
<tr>
<td>nonparametric</td>
<td>infinite</td>
<td>local</td>
<td>robust</td>
</tr>
</tbody>
</table>
Introduction

Part II: Learning Theory for Supervised Learning

Loss, Risk, and Optimal Learner

Learning Process: ERM

Restricted Model Class $\mathcal{F}$

---

Graphs showing different fits:
- **Linear fit**
- **2-order polynomial fit**
- **3-order polynomial fit**
- **Smoothing spline (lambda=0.85)**
Let $\mathcal{F}$ denote the restricted model space.

- Denote $f^* = \arg \min EPE(f)$, the optimal learner. The minimization is taken over all measurable functions.
- Denote $\tilde{f} = \arg \min_{\mathcal{F}} EPE(f)$, the best learner in the restricted space $\mathcal{F}$.
- Denote $\hat{f} = \arg \min_{\mathcal{F}} R_{\text{emp}}(f)$, the learner based on limited data in the restricted space $\mathcal{F}$.

\[
EPE(\hat{f}) - EPE(f^*) = [EPE(\hat{f}) - EPE(\tilde{f})] + [EPE(\tilde{f}) - EPE(f^*)] = \text{estimation error} + \text{approximation error}.
\]
The gap $\text{EPE}(\hat{f}) - \text{EPE}(\tilde{f})$ is called the \textit{estimation} error, which is due to the limited training data.

The gap $\text{EPE}(\tilde{f}) - \text{EPE}(f^*)$ is called the \textit{approximation} error. It is incurred from approximating $f^*$ with a restricted model space or via regularization, since it is possible that $f^* \notin \mathcal{F}$.

This decomposition is similar to the bias-variance trade-off, with

- estimation error playing the role of variance
- approximation error playing the role of bias.