SI151 Discussion 2

Jiachun Jin, jinjch@shanghaitech.edu.cn

Review

1. Some Tricky Points

- Overview of supervised learning
  - Statistical decision theory
    1. The general idea: once given a metric to measure the effectiveness of a learned model, what is the theoretically optimal predictor?
       1. l_2 loss in regression, the regression function: $E[Y|X=x]$
       2. 0-1 loss in classification, the Bayesian classifier: $\underset{g\in G}{\operatorname{argmax}} \Pr(g \mid X=x)$
    2. Note that $EPE(f)$ is a function of function, we want to search for a group of functions such that: $\hat{f} = \arg\min_f EPE(f)$
       - solution 1: calculus of variations 🤔😵
       - solution 2: minimize EPE pointwise

$$
\begin{aligned}
\hat{f}(X=x) &= \arg\min_c E_{Y|X}[(Y|X-c)^2|X=x] \\
&= E_{Y|X}[Y|X=x]
\end{aligned}
$$

- Curse of dimensionality

- The bias-variance decomposition

$$
\begin{aligned}
\mathrm{MSE}(x_0) &= E_{\mathcal{T}}[f(x_0)-\hat{y}_0]^2 \\
&= E_{\mathcal{T}}[\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0) + E_{\mathcal{T}}(\hat{y}_0) - f(x_0)]^2 \\
&= E_{\mathcal{T}}\left[(\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0))^2 + 2(\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0))(E_{\mathcal{T}}(\hat{y}_0) - f(x_0)) + (E_{\mathcal{T}}(\hat{y}_0) - f(x_0))^2\right] \\
&= E_{\mathcal{T}}\left[(\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0))^2\right] + (E_{\mathcal{T}}(\hat{y}_0) - f(x_0))^2 \\
&= \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + \mathrm{Bias}^2(\hat{y}_0)
\end{aligned}
$$

1. $\hat{y}_0 := y(x_0;\mathcal{T})$: we use the training set $\mathcal{T}$ to learn a model, and then use the learned model to do prediction over a test point x_0, the predicted label is denoted as $\hat{y}_0$, different training set can produce different learned model, as well as different predicted label for x_0, so we compute the average with expectation.

2. What is a distribution of the training set?

2. Bayesian Learning and Non-Bayesian Learning

- non-probabilistic model

- probabilistic model with parameter θ, training dataset $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$, a test point x_0

  1. take θ as unknown constants

  - the learning phase: to solve an optimization problem

  $$\hat{\theta} = \arg\min_\theta \mathcal{L}(\mathcal{D}, \theta) + \Omega(\theta)$$

  the prediction phase: directly plug the learned $\hat{\theta}$ into the model and do prediction to x_0

  2. take θ as unknown random variables (Bayesian Learning), y_0: a random variable denotes the label of the test point x_0

  - the learning phase: to do probability inference, as well as solving an integral problem

$$
\begin{aligned}
P(\theta|\mathcal{D}) &= \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} \\
&= \frac{P(\mathcal{D}|\theta)P(\theta)}{\int P(\mathcal{D}|\theta)P(\theta)d\theta} \\
&= \frac{P(\mathcal{D}|\theta=k)P(\theta=k)}{\sum_k P(\mathcal{D}|\theta=k)P(\theta=k)}
\end{aligned}
$$

- the prediction phase, we want the posterior distribution of y_0 after observing the training dataset $\mathcal{D}$:

  $$P(y_0|\mathcal{D}) = \int_\theta P(y_0|\theta)P(\theta|\mathcal{D})d\theta$$

  Then we use statistical decision theory to do decision with the above distribution.

  > remark1: In a purely Bayesian learning setting, learning is the same as inference.
  >
  > remark2: Give a measure of the the confidence of prediction, also can do sequential learning.
  >
  > remark3: We will not cover a lot of Bayesian learning in this course.

- What about MAP?
  - choose the mode of the parameter's posterior distribution

  $\hat{\theta} = \arg\max P(\theta|\mathcal{D})P(\theta)$

- An example

  > 拉普拉斯的太阳问题: 过去10000天，我们都观测到太阳照常升起，明天太阳照常升起的概率是多少?

The Gauss-Markov Theorem

Theorem: *The least squares estimator has the lowest sampling variance within the class of linear unbiased estimators.*

Proof:

Bayesian Linear Regression

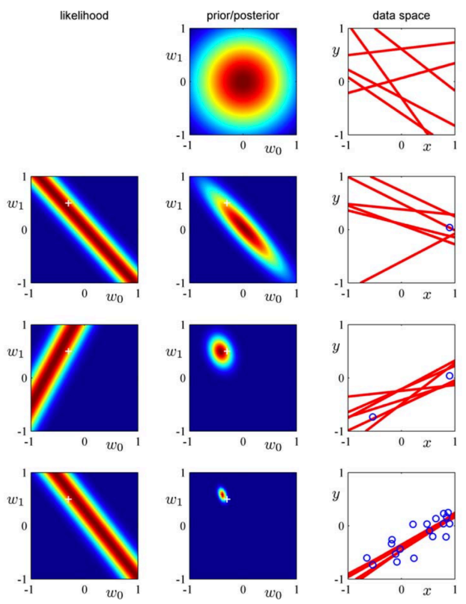
1. Parameter Distribution

- $y = \mathrm{x}^\top w + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$
- prior of w: $w \sim \mathcal{N}(0, \alpha^{-1} I_d)$,

2. An Example

- $y = w_1 x + w_0$
- ground-truth: $w_0 = -0.3, w_1 = 0.5$, data points are sampled from $y = 0.5x - 0.3$ with an additive noise.
- The last posterior is used as the next prior.
- How to make prediction?

$$p(y|\text{Data}, x_0) = \int p(y|w, x_0)p(w|\text{Data})dw$$



An Incomplete Proof of SVD(tentative)

Theorem: ["thin SVD"] $A \in \mathbb{R}^{m\times n}$, *rank(A) = r*, $\exists$ *orthonormal basis* $U_A \in \mathbb{R}^{m\times r}$ of $\mathcal{R}(A)$ and orthonormal basis $V_A \in \mathbb{R}^{n\times r}$ * of $\mathcal{R}(A^\top)$, such that $A = U_A \Sigma_A V_A^\top$, where $\Sigma_A = \begin{bmatrix} \Lambda_1^{1/2} & & \\ & \ddots & \\ & & \Lambda_k^{1/2} \end{bmatrix}$ with $\Lambda_i = \lambda_i I_{g_i}$, λ_i is eigenvalue of $AA^\top$, and λ_k is the smallest non-zero eigenvalue of $AA^\top$, the set of $AA^\top$'s eigenvalues: $\{\lambda_1 > \ldots > \lambda_s\}$, and $k = s$ or $k = s - 1$

Proof: