INSEA

39

Machine learning: Frequentist vs Bayesian

Frequentist machine learning

Bayesian machine learning

Fonction de prédiction f_θ paramétrée par $\theta \in \mathbb{R}^p$.

On obtient M fonctions de prédiction $(f_{\theta_1}, \ldots, f_{\theta_M})$

On optimise:

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \text{loss}(f_\theta(\mathbf{x}_i), y_i) \quad + \quad \frac{1}{C} \, \text{pénalité}(\theta)$$

On obtient une fonction de prédiction optimale $f_{\theta^\star}$

Choisir le meilleur C par validation croisée

On obtient une prédiction moyenne avec un intervalle de crédibilité

Pour chaque $\mathbf{x}_i$ on a M prédictions: une distribution de prédictions

On simule une MCMC $\theta_1, \ldots, \theta_M \sim$ loi a posteriori $\theta | y_i, \mathbf{x}_i$

C est simulé suivant un modèle hiérarchique avec $C \sim$ hyperprior

θ est un vecteur aléatoire suivant une loi a priori π de variance $\propto C$

1. Facile à expliquer et à implémenter

2. Adapté pour des quantités de données gigantesques (Optimisation distribuée, stochastique)

3. Optimisation (souvent) non-convexe: dépend de l'initialisation

1. Quantifie l'incertitude des prédictions: essentiel pour des applications sensibles (diagnostique médical, voitures autonomes …)

2. Basé sur la simulation MCMC (lente en grande dimension / risque de divergence)

Bayesian machine learning

Frequentist machine learning

Fonction de prédiction f_θ paramétrée par $\theta \in \mathbb{R}^p$.

On optimise:

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \text{loss}(f_\theta(\mathbf{x}_i), y_i) \quad + \quad \frac{1}{C} \,\text{pénalité}(\theta)$$

On obtient une fonction de prédiction optimale $f_{\theta^\star}$

Choisir le meilleur C par validation croisée

1. Facile à expliquer et à implémenter
2. Adapté pour des quantités de données gigantesques (Optimisation distribuée, stochastique)
3. Optimisation (souvent) non-convexe: dépend de l'initialisation

Bayesian machine learning

θ est un vecteur aléatoire suivant une loi a priori π de variance $\propto C$

On simule une MCMC $\theta_1, \ldots, \theta_M \sim$ loi a posteriori $\theta | y_i, \mathbf{x}_i$

On obtient M fonctions de prédiction $(f_{\theta_1}, \ldots, f_{\theta_M})$

Pour chaque $\mathbf{x}_i$ on a M prédictions: une distribution de prédictions

On obtient une prédiction moyenne avec un intervalle de crédibilité

C est simulé suivant un modèle hiérarchique avec $C \sim$ hyperprior

1. Quantifie l'incertitude des prédictions: essentiel pour des applications sensibles (diagnostique médical, voitures autonomes …)
2. Basé sur la simulation MCMC (lente en grande dimension / risque de divergence)

INSEA

Best of both worlds: