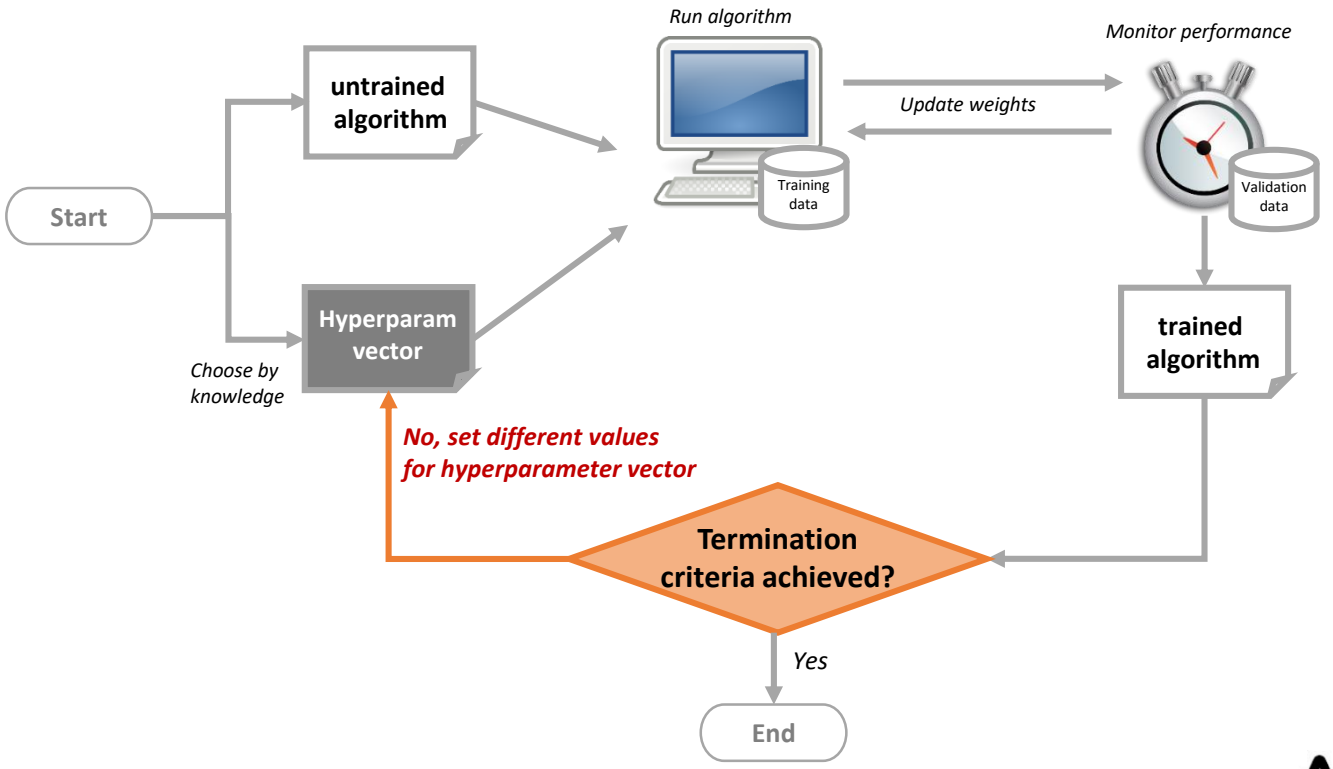
TA Session 3-2

2018.10.18

Wonjong Rhee, Hyunghun Cho, Daeyoung Choi
Seoul National University
Graduate School of Convergence Science and Technology
Applied Data Science Lab.

HPO for Traditional ML

- Training with different values of hyperparameter vector is **manually** repeated until termination criteria achieved
- More efficient than automatic solution



Task 1: Grid search CV

- Carseats datasets를 변형해서 Sales 가 8보다 높을지를 예측하는 Random Forest (RF) 모델을 학습하세요.
  - [Hint]: *ISLR/Chapter 8.ipynb*
  - Training set / test set 은 7:3 으로 분리
  - RF 의 hyperparameters 와 값은 임의로(by your prior knowledge) 선정. (Refer to documentation)
  - 예측 결과를 Precision/Recall, F1-score 및 confusion matrix로 분석

- 두 개의 hyperparameters 와 grid 에 포함될 값들을 선정하여 Grid search CV 함수를 수행하고 결과를 출력하세요.
  - [Hint]: *ISLR/Chapter 8.ipynb*

- hyperparameter pair 를 x, y 축으로, mean test score를 z축으로 하는 3D wireframe plot를 그리고 분석하세요.
  - hyperparameter 별 성능 영향 분석

Task 2: Randomized search CV

- Task 1의 RF 모델을 이용해 randomized search CV 함수를 수행하세요.
  - hyperparameter 별 range setting는 scipy.stats 참조
    - e.g. uniform random
      - for integer: scipy.stats.randint
      - for float: scipy.stats.uniform
    - 그 외 distribution 는 scipy distributions 참고

- 앞서 선택한 두 개의 hyperparameters 와 range 에 min, max 값들로 Randomized search CV 함수를 수행하고 결과를 출력하세요.

- hyperparameter pair 를 x, y 축으로, mean test score를 z축으로 하는 3D scatter plot를 그리고 분석하세요.
  - Task 1의 plot 결과와 비교

Task 3: CV with SVM (optional)

- Task 1,2 를 SVM 함수를 이용해 Grid search CV, Randomized search CV를 수행하세요.
  - [hint]: *ISLR/Chapter 8.ipynb*
  - **[caveat]** SVM은 hyperparameter에 따라 수행이 끝나지 않거나 매우 오래 걸릴 수 있습니다.

Sequential Modeling Algorithm

- Sequential Model-Based Optimization

SMBO(f, M_0, T, S)

1 $\mathcal{H} \leftarrow \emptyset$,
2 For $t \leftarrow 1$ **to** T,
3 $x^* \leftarrow \text{argmin}_x \; S(x, M_{t-1})$,
4 Evaluate $f(x^*)$, ▷ *Expensive step*
5 $\mathcal{H} \leftarrow \mathcal{H} \cup (x^*, f(x^*))$,
6 Fit a new model M_t to $\mathcal{H}$.
7 **return** $\mathcal{H}$

Figure 1: The pseudo-code of generic Sequential Model-Based Optimization.

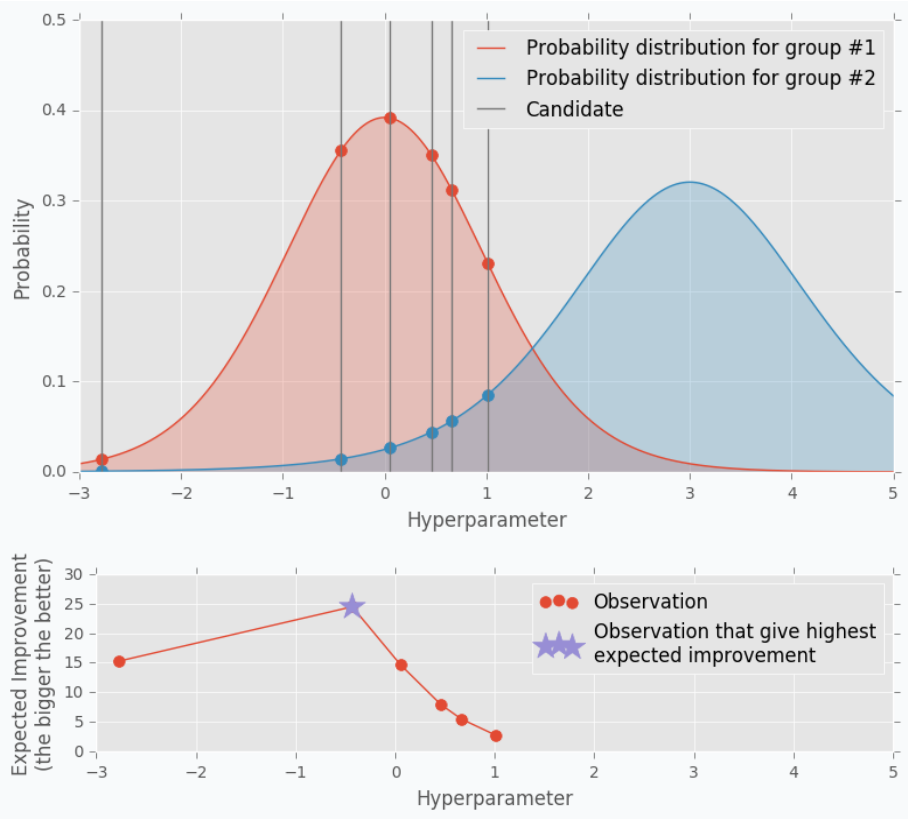
- Bayesian Optimization (BO)
  - developing a statistical model of the function from hyperparameter values to the objective evaluated on a validation set
  - exploration vs exploitation tradeoff

※ Source: F. Diehl and A. Jauch

TPE

- Create two hierarchical processes $\ell(\mathbf{x})$, $g(\mathbf{x})$ acting as generative models for all domain variables
- Model the domain variables when the object function is below and above a specified quantile y^*,

$$p(\mathbf{x}|y,\mathcal{D}) = \begin{cases} \ell(\mathbf{x}), & \text{if } y < y^*, \\ g(\mathbf{x}), & \text{if } y \geq y^*. \end{cases}$$

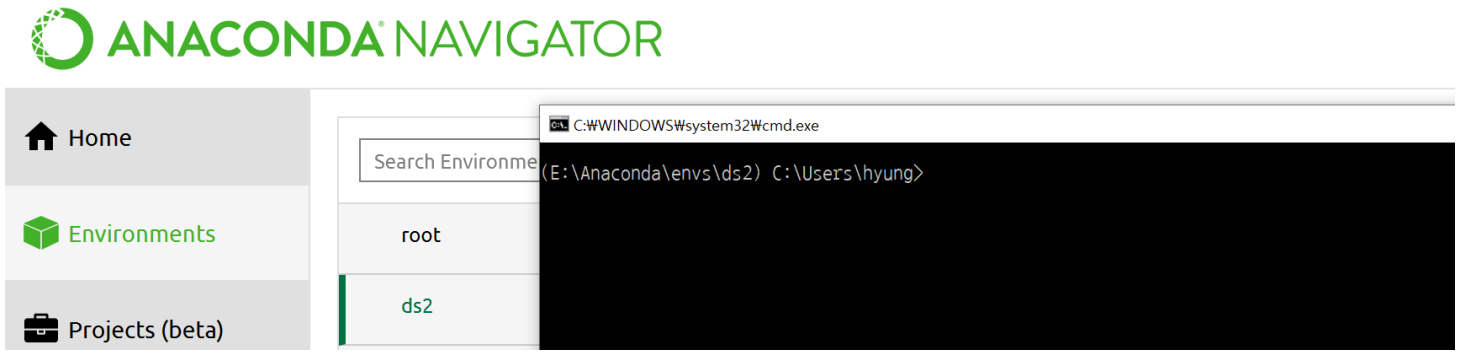


$$EI(\mathrm{x}) = \frac{\ell(\mathbf{x})}{g(\mathbf{x})}$$

Source: NeuPy

Tutorial: Bayesian Optimization

- hyperopt 설치
  - Anaconda navigator 에서 개발 환경을 설정한 virtualenv 를 선택한 후 ► 버튼에 오른쪽 클릭하고 “Open Terminal” 선택



  - cmd 창에서 ‘pip install hyperopt’ 입력하여 설치

- 도움말 참조

유의사항

- Task는 구현 자체보다는 결과에 대한 분석과 이해를 중심으로 학습 하시기 바랍니다.
- 제공된 2018_fall_session_3_task_HPO.ipynb 파일에 코드를 작성하고 설명을 단 뒤 ipynb 파일로 제출해주세요.
- 제출 기한은 진행에 따라 정하겠습니다.
- 다음의 이메일로 choid@snu.ac.kr (최대영 조교) 로 제출하시고 보내실 때 반드시 조별 대표자 성명을 기록해주시기 바랍니다.