基本概念

数据

- feature, target(label)
- training, test set
- validation set

步骤

- *feature extraction:* 从输入数据里抽feature，从而把问题变容易，或让计算变快
- *training*:学习feature到target的映射，通常涉及基于某个模型假设解一个优化问题
- *model selection*: 从各种模型中找一个最好的
- *testing*: 评价性能

问题分类

- supervised: 分类、回归
- unsupervised: 聚类、密度估计、 visualization
- reinforcement: 从经验中学习

Regularization

- overfitting: 模型过于复杂，导致能拟合任意training set
- regularization: 通常加一个penalty term来discourage

理论基础

- Bayesian theory: ML, MAP, full Bayesian什么的
- decision theory: LDA什么的，一般对regression error或者误分类有一个cost，最小化cost的期望。
- Information theory: Boltzmann machine之类的

概率论基础

PDF（连续）/PMF(离散）：$p(x)$　　CDF: $P(x) = \int p(x)\,dx$

分布转换

$$p(X) = \int_Y p(X, Y) \qquad p(X, Y) = p(Y|X)p(X)$$

统计量

mean: $E[x] = \int p(x) \cdot x\,dx$　　variance: $E[x] = \int p(x) \cdot x^2\,dx$

sample mean: $\mu_{ML} = \frac{1}{N}\sum_n x_n$　　sample variance: $\sigma_{ML} = \frac{1}{N}\sum_n x_n^2$

样本均值的数学期望不见得是mean, variance

信息论基础

熵: $H[x] = -\int p(x)\, ln\, p(x)\, dx$

条件熵: $H[y|x] = -\iint p(y, x)\, ln\, p(y|x)\, dy\, dx = H[x, y] - H[x]$

相对熵 (KL divergence): $KL(p|q) = -\int p(x)\, ln\{\frac{q(x)}{p(x)}\} dx \geq 0$

互信息 $I(x, y) = KL(p(x, y)|p(x)p(y)) = H[x] - H[x|y] = H[y] - H[y|x]$

Bayesian inference基础

training方法分类

- 最大似然(ML): $\theta_{ML} = argmax\, p(X|\theta)$
- 最大后验(MAP): $\theta_{MAP}(\alpha) = argmax\, p(\theta|X) = argmax\, p(X|\theta)p(\theta|\alpha)$
- 全 Bayesian: computes the posterio distribution $p(\theta|X)$ itself

model selection方法分类

- 基于valiation的方法：单独抽一个和training/testing独立的validation set出来用于比较不同模型的generalization power
    - cross-validation: 把training data分为N份，S-1份用来training, S份用来validate
    - leave-on-one: 用于training数据特别少的情况，循环S次，每次取其中一分作validation，最后取model中总的validation error最小的
- 基于信息论的方法
    - Akaike Information Criterion(AIC): $ln\, p(X|\theta) - M$，其中M是维数
    - BIC: TODO

The curse of dimensionality: 模型维度升高时，出现了很多新的难以克服的问题

- 模型在高维空间所占体积巨大，以至于不容易得到充分采样
- 低维空间的直觉对高维不成立