Basic concepts

Data

- feature, target vectors
- training, test set
- validation set

General steps

- ***feature extraction:*** training sample -> feature
- ***training***: learn feature-to-target mapping
  often by minimizing an error function
- ***model selection***: choose model with best performance on validation set
- ***testing***: evaluate performance

Learning problems

- **supervised:** classification, regression
- **unsupervised:** clustering, density estimation, visualization
- **reinforcement:** tradeoff between exploration and exploition

Regularization

- **overfitting:** the model obtained from the learning phase can fit training set perfectly, but predicts test data poorly
- **regularization:** add a penalty term to the error function to offset model complexity

Three views to perceive a learning problem

- **Bayesian theory:** maximize posteria probability
- **Decision theory:** define a cost function for each error, and minimize expected cost over model parameter (usually the decision boundary/surfaces of a linear model)
- **Information theory**: minimize model entrophy

Probability 101

Distribution functions:

- probabilisty density(continuous r.v.)/probabilisty mass(discrete r.v.) fucntion: $p(x)$
- cumulative distribution function: $P(x) = \int p(x)\,dx$

Rules

- joint-to-marginal distribution: $p(X) = \sum_Y p(X, Y)$
- conditional-to-joint distribution: $p(X, Y) = p(Y|X)p(X)$

Statistics

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: $\sigma_{ML} = \frac{1}{N} \sum_n x_n^2$

Note the mathematical expectation of sample mean and variance may not equal to true model mean and variance.

Information theory 101

Entropy

- marginal entropy: $H[x] = -\int p(x) \, ln \, p(x) \, dx$
- conditional entropy: $H[y|x] = -\iint p(y, x) \, ln \, p(y|x) \, dy \, dx = H[x, y] - H[x]$

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

Mutual information $I(x, y) = KL(p(x, y)|p(x)p(y)) = H[x] - H[x|y] = H[y] - H[y|x]$

More on Bayesian Approach

Learning approach

- **Maximum likelihood(ML):** $\theta_{ML} = argmax \, p(X|\theta)$
- **Maximum posterior(MAP):** $\theta_{MAP}(\alpha) = argmax \, p(\theta|X) = argmax \, p(X|\theta)p(\theta|\alpha)$
- **Full Bayesian:** computes the posterio distribution $p(\theta|X)$ itself

Model selection by validation

- **validation:** test learned model on a third set separate from training/testing
- **cross-validation:** partition data into S parts, use it for validation, use the other S-1 parts for training
- **leave-on-one:** repeat cross-validation for each of the S parts
  (useful for scarse data, but need to train S times)

Model selection by information criteria

- **Akaike Information Criterion(AIC):** $ln \, p(X|\theta) - M$, where M is model dimensionality
- BIC: TODO

The curse of dimensionality: with increase of independent parameters, the volume of data a model can represent signifantly increases. Therefore increasing the model complexity too much will reduce its generalization power