Fundamentals of Machine Learning

BAYESIAN DECISION THEORY

Acknowledgments: Adapted from slides at https://probml.github.io/pml-book/teaching1.html by Prof. Saw Shier Nee

Outline

- Bayesian Decision Theory
- Evaluation methods
- Hypothesis Testing

Decision Theory

We assume the decision maker, or agent, has a set of possible actions, A, to choose from.

Every action has cost and benefits depending on underlying state of nature $h \in \mathcal{H}$.

Encode this information to loss function $\ell(h, a)$, loss which we incur if action $a \in A$ is taken at state of nature, $h \in \mathcal{H}$.

Risk: $$R(a|\boldsymbol{x}) \triangleq \mathbb{E}_{p(h|\boldsymbol{x})}[\ell(h,a)] = \sum_{h \in \mathcal{H}} \ell(h,a) p(h|\boldsymbol{x})$$

Optimal Policy: $$\pi^*(\boldsymbol{x}) = \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{E}_{p(h|\boldsymbol{x})}[\ell(h,a)]$$

Utility Function

Maximum expected utility: $$\pi^*(\boldsymbol{x}) = \operatorname*{argmax}_{a \in \mathcal{A}} \mathbb{E}_h[U(h,a)]$$

$$U(h,a) = -\ell(h,a)$$

Classification- Accuracy

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	**True positive** (TP), hit	**False negative** (FN), type II error, miss, underestimation
	Negative (N)	**False positive** (FP), type I error, false alarm, overestimation	**True negative** (TN), correct rejection

Be aware of class imbalance.

Accuracy = (TP + TN) / (TP + FN + FP + TN)

True Positive Rate (TPR) e.g., predict how many disease correctly in disease cohort

TPR = TP / (TP + FN)

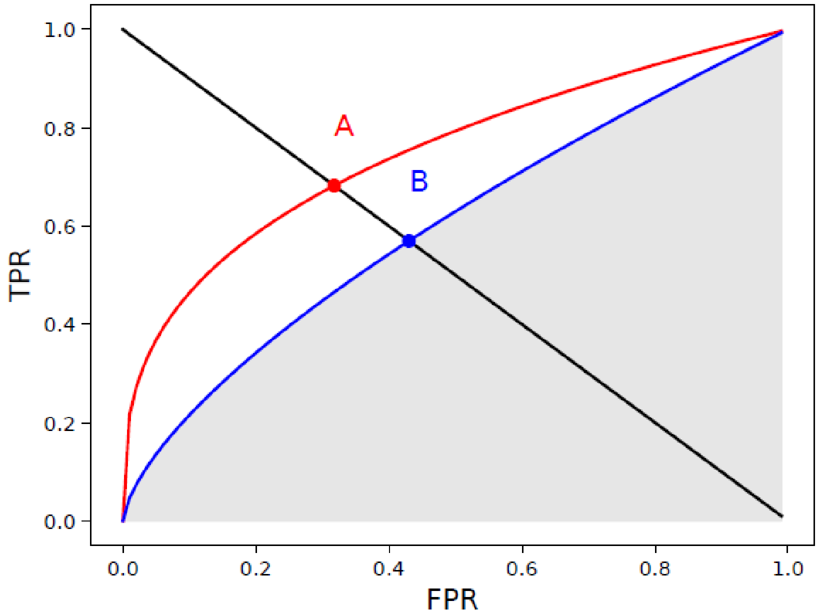
False Positive Rate (FPR) e.g., predict how many disease in healthy cohort

FPR = FP / (FP + TN)

Classification- ROC curve

Area under ROC curve is computed.

The higher the AUROC, the better the classifier.

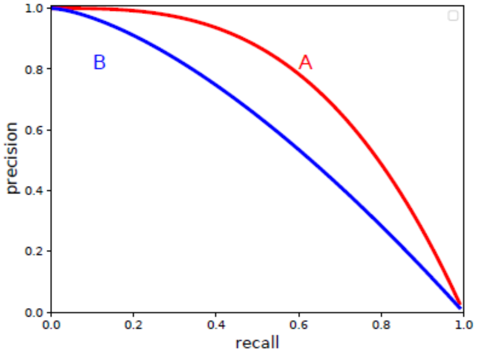


		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection

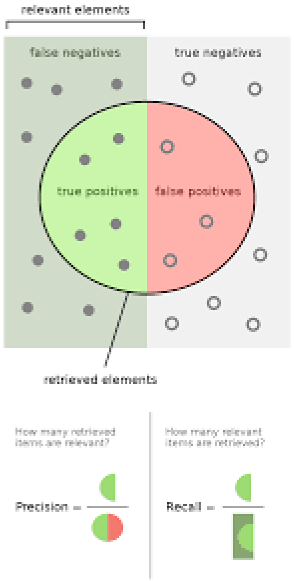
EER: FPR = FNR = 1 - TPR

TPR = TP / (TP + FN), e.g., predict how many disease correctly in disease cohort
FPR = FP / (FP + TN), e.g., predict how many disease in healthy cohort

Classification- Precision-Recall Curve



Recall = TPR = TP / (TP + FN) - predict how many positive correctly in positive cohort
Precision = Positive Predictive Value (PPV) = TP / (TP +FP) - predict how many positive correctly if prediction is positive.



Area under PR curve is computed.

The higher the AUC of PR curve, the better the classifier.

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	**True positive** (TP), hit	**False negative** (FN), type II error, miss, underestimation
	Negative (N)	**False positive** (FP), type I error, false alarm, overestimation	**True negative** (TN), correct rejection

Hypothesis Testing

Suppose we have two hypotheses or models, commonly called the null hypothesis, M_0, and the alternative hypothesis, M_1, and we want to know which one is more likely to be true.

The optimal decision to pick alternative hypothesis iff

$$P(M_1|D) > P(M_0|D)$$

$$\frac{P(M_1|D)}{P(M_0|D)} > 1$$

Hypothesis Testing

If we use uniform prior $p(M_0) = p(M_1) = 0.5$, the decision rule becomes

$$\frac{P(D \mid M_1)}{P(D \mid M_0)} > 1$$

Bayes Factor = ratio of marginal likelihood of two models

$$B_{1,0} \triangleq \frac{P(D \mid M_1)}{P(D \mid M_0)}$$

Bayes factor $BF(1,0)$	Interpretation
$BF < \frac{1}{100}$	Decisive evidence for M_0
$BF < \frac{1}{10}$	Strong evidence for M_0
$\frac{1}{10} < BF < \frac{1}{3}$	Moderate evidence for M_0
$\frac{1}{3} < BF < 1$	Weak evidence for M_0
$1 < BF < 3$	Weak evidence for M_1
$3 < BF < 10$	Moderate evidence for M_1
$BF > 10$	Strong evidence for M_1
$BF > 100$	Decisive evidence for M_1

Table 5.6: Jeffreys scale of evidence for interpreting Bayes factors.

Bayesian Model Selection

Suppose we have a set *M* of more than 2 models, and we want to pick the most likely.

The optimal action is picking the most probable model

$$\hat{m} = \underset{m \in \mathcal{M}}{\operatorname{argmax}}\, p(m|\mathcal{D})$$

where

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m \in \mathcal{M}} p(\mathcal{D}|m)p(m)}$$

Bayesian Model Selection

If the prior over models is uniform, p(m) = 1/|M|, then the Maximum A Posterior (MAP) is given by

$$\hat{m} = \underset{m \in \mathcal{M}}{\operatorname{argmax}}\, p(\mathcal{D}|m)$$

The quantity $p(\mathcal{D}|m)$ is given by

prior

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta}, m)p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$$

This is known as the **marginal likelihood**, or the **evidence** for model m.

Occam's razor

Consider two models, a simple one, m_1, and a more complex one, m_2 .

Suppose that both can explain the data by suitably optimizing their parameters.

Intuitively we should prefer m_1, since it is simpler and just as good as m_2.

This principle is known as Occam's razor

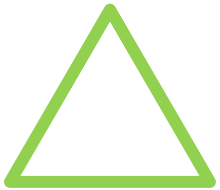
Bayesian Occam’s razor effect

Intuition from Squid Game

Complex Model

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\theta, m)p(\theta|m)d\theta$$

Many θ averaged over the parameter space



Simple Model

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\theta, m)p(\theta|m)d\theta$$

Few θ averaged over the parameter space