Tutorial 2

Key notes in lecture slides: 2-statistical learning (2.4 ~ 2.6)

2.4 Evaluation of Classifiers

Evaluating classification methods

北师香港浸会大学
BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY

- Predictive accuracy:

$$Accuracy = \frac{Number\ of\ correct\ classifications}{Total\ Number\ of\ test\ cases}$$

- Should not be used when class labels are imbalanced.
  - E.g. class 0: 99%, class 1: 1%. Then a model that simply predicts 0 for any input has an 99% accuracy.
  - Better to use other metrics such as recall, precision, F1-score.

Evaluating classification methods

北师香港浸会大学
BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY

- **Holdout set**: The available data set D is divided into two disjoint subsets,
  - the training set D_{train} (for learning a model)
  - the test set D_{test} (for testing the model)
- **Important**: training set should not be used in testing and the test set should not be used in learning.
  - Unseen test set provides a **unbiased estimate** of accuracy.
    - I.e. $\mathbb{E}$[test accuracy] = Population accuracy.
    - Thus, test accuracy is a good measure of how model would perform on population, even though we do not report $\mathbb{E}$[test accuracy] in practice.
    - We can compute 95% confidence interval to indicate the population accuracy using test accuracy.
- The test set is also called the holdout set. (the examples in the original data set D are all labeled with classes.)
- This method is mainly used when the data set D is large.

Precision and recall measures

- Used in information retrieval and text classification.
- We use a confusion matrix to introduce them.

	Classified Positive	Classified Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

where

TP: the number of correct classifications of the positive examples (**true positive**),
FN: the number of incorrect classifications of positive examples (**false negative**),
FP: the number of incorrect classifications of negative examples (**false positive**), and
TN: the number of correct classifications of negative examples (**true negative**).

Precision and recall measures (cont…)

	Classified Positive	Classified Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

$$p = \frac{TP}{TP + FP}. \qquad r = \frac{TP}{TP + FN}.$$

- Precision *p* is the number of correctly classified positive examples divided by the total number of examples that are classified as positive.
- Recall *r* is the number of correctly classified positive examples divided by the total number of actual positive examples in the test set.

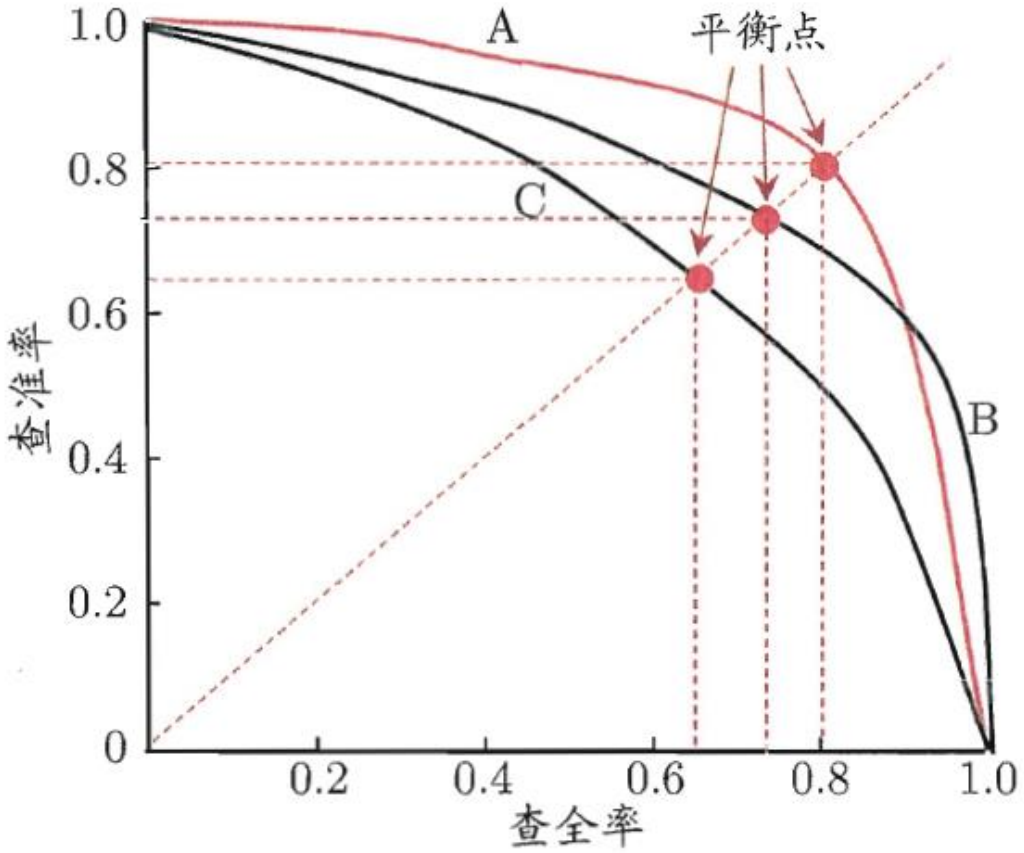
An example

	Classified Positive	Classified Negative
Actual Positive	1	99
Actual Negative	0	1000

- This confusion matrix gives
  - precision p = 100% and
  - recall r = 1%

  because we only classified one positive example correctly and no negative examples wrongly.
- Note: precision and recall only measure classification on the positive class.

P-R Curve



1. According to the prediction results of the learner, the higher the ranking, the more likely it is to be a positive example.
2. If the samples are taken as positive examples one by one in this order, a set of precision and recall can be achieved.

Question?

Which one is better?

F1-value (also called F1-score)

北师香港浸会大学
BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY

- It is hard to compare two classifiers using two measures. F_1 score combines precision and recall into one measure

$$F_1 = \frac{2pr}{p+r}$$

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

F_1-score is the harmonic mean of precision and recall.

$$F_1 = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$

- The harmonic mean of two numbers tends to be closer to the smaller of the two.
- For F_1-value to be large, both *p* and *r* must be large.

Which one should we use, P or R?

北师香港浸会大学
BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY

- Classification: generate a real value or probability prediction for the test sample, and then compare this prediction value with a classification threshold. If it is greater than the threshold, it will be classified as positive, otherwise negative.
  - Equivalent to sorting, "most likely" is the positive case at the front, and "least likely" is the positive case at the back. The classification process is equivalent to dividing the sample into two parts by a certain "cut-off point" in this sorting, the former part is judged as positive cases, and the latter part is judged as negative cases.
  - If more attention is paid to P, the front position will be truncated, such as product recommendation.
  - If more attention is paid to R, the back position will be truncated, such as fugitive search.

Receive operating characteristics curve

北师香港浸会大学
BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY

- It is commonly called the ROC curve.
- It is a plot of the true positive rate (TPR) against the false positive rate (FPR).
- True positive rate (i.e., recall):

$$TPR = \frac{TP}{TP + FN} \qquad \frac{correctly\ classified\ positve}{All\ Positive}$$

- False positive rate (假阳):

$$FPR = \frac{FP}{TN + FP} \qquad \frac{negative\ that\ are\ incorrectly\ treated\ as\ positve}{All\ Negative}$$

Example ROC curves

ROC curve: TPR against FPR under different **classification threshold.**

- Suppose we have a classification model, it classifies samples with probability greater than a threshold as positive and the remaining as negative.
- Then we can compute TPR and FPR for this threshold value and plot a (FPR, TPR) point.
- We choose threshold values range from 1 to 0 and plot (FPR, TPR) for each of them, then we will get the ROC curve, shown in Figure 3.8 on the right.

- In general, threshold decreases from 1 to 0 from left of the ROC curve to the right.

- Please read:
  - https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

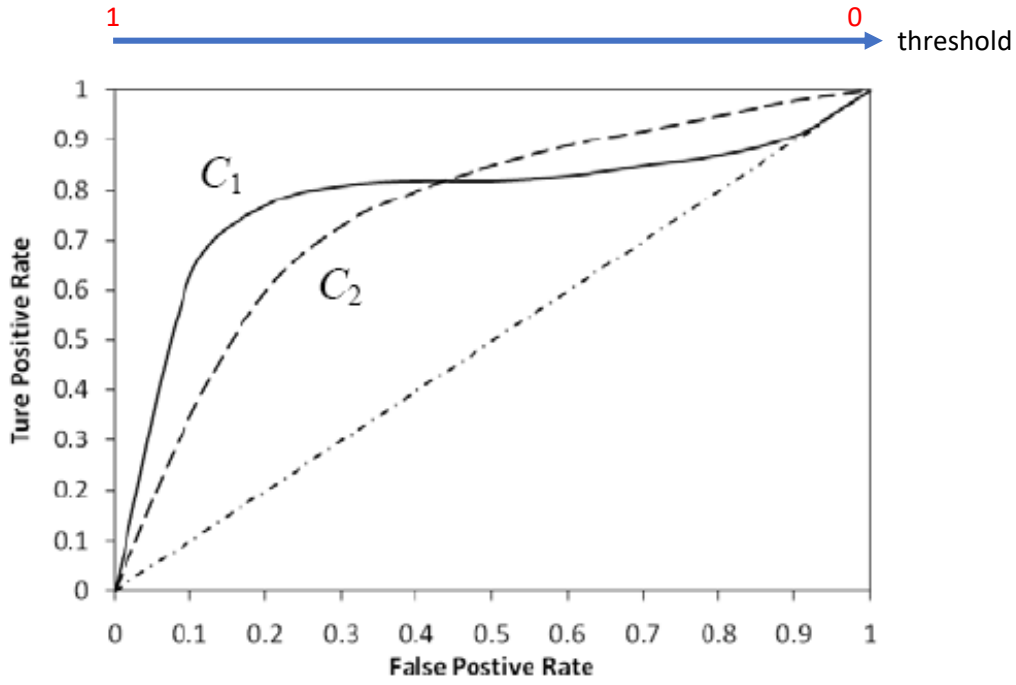


Fig. 3.8. ROC curves for two classifiers (C_1 and C_2) on the same data

Area under the curve (AUC)

北师香港浸会大学
BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY

- Which classifier is better, C_1 or C_2?
  - It depends on which region you talk about.

- Can we have one measure?
  - Yes, we compute the area under the curve (AUC),

- If AUC for C_i is greater than that of C_j, in general, we can say that C_i is better than C_j.
  - A model with greater AUC means that it is generally better than a model with less AUC over all classification thresholds.
  - If a classifier is perfect, its AUC value is 1.
  - If a classifier makes **all random guesses**, its AUC value is 0.5.
  - We may choose models with **smaller** AUC in real world applications.

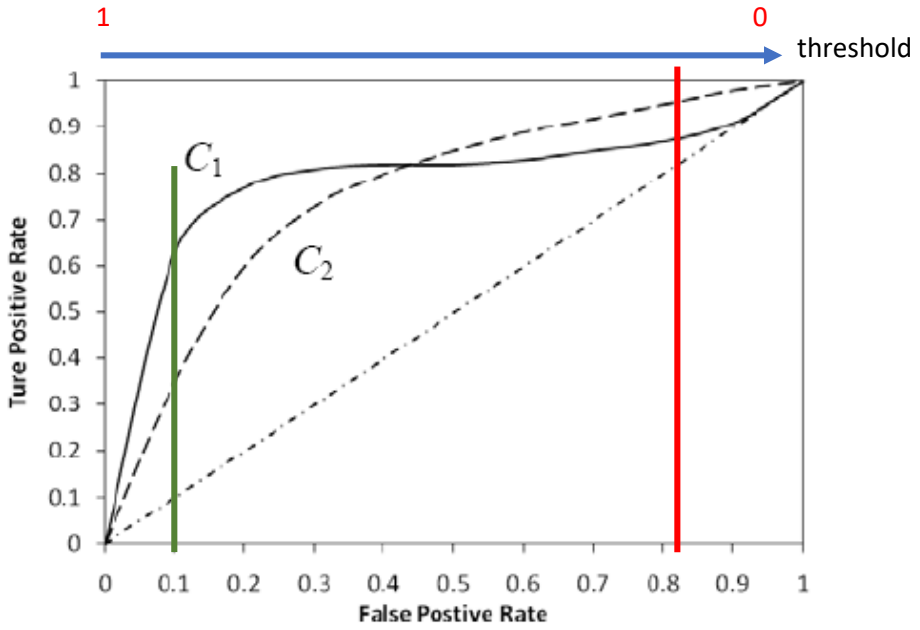
Area under the curve (AUC)

北师香港浸会大学
BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY

- If AUC for C_i is greater than that of C_j, in general, we can say that C_i is better than C_j.
  - We may prefer models with **smaller** AUC in real world applications.
  - Case 1:
    - When detecting diseases, we don't want to miss any potential patient. So we can accept as many patients to be correctly spotted at the cost of some healthy people to be wrongly classified as ill. In this case, c2 is a better model.
  - Case 2:
    - When detecting spam email, we don't want a normal email to be mis-classified as an spam.
    - In this case, c1 is a better model.

Case 1: **c2** is a better model

Case 2: **c1** is a better model



Likelihood Function (似然函数)

北师香港浸会大学
BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY

- The likelihood function is a function of the parameter w about the statistical model $p(x; w)$
  - Probability $p(x; w)$ describes the distribution of the random variable x when the parameter w is fixed.
  - Likelihood $p(x; w)$ describes the influence of different parameters w on the distribution of a known random variable x.

- We can also denote the likelihood function as $\mathcal{L}(w|x)$ or $\mathcal{L}(\theta|x)$
  - Note that $\mathcal{L}(w|x) = P(x\ |w)$

Maximum Likelihood Estimation (MLE)

北师香港浸会大学
BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY

- Assume $X_1, X_2, \cdots, X_n$ are samples from X, the probability of the observed samples occurrence is $\prod_{i=1}^{n} p(x_i; w)$.
- Likelihood function:

$$L(w) = L(x_1, x_2, \cdots, x_n; w) = \prod_{i=1}^{n} p(x_i; w)$$

- Maximize the likelihood function to get $\hat{w}$:

$$\hat{w} = \text{argmax}_w L(x_1, x_2, \cdots, x_n; w)$$

where $\hat{w}(x_1, x_2, \cdots, x_n)$ is maximum likelihood estimation and $\hat{w}(X_1, X_2, \cdots, X_n)$ is maximum likelihood statistic.

- Log Likelihood Equation:

$$dL(w)/dw = 0 \rightarrow d\log L(w)/dw = 0$$

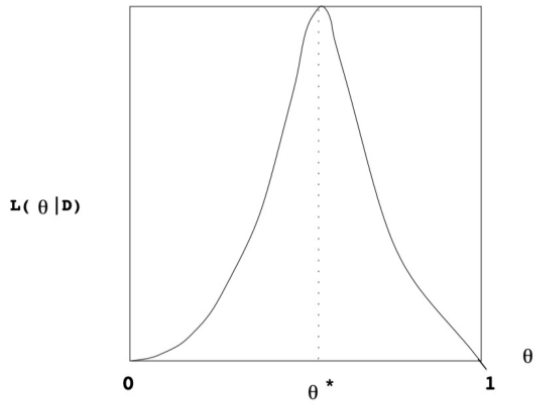


Figure 2: Likelihood function for the sequence $\mathbf{D} = \{H, T, H, T, T, H\}$

Likelihood Function in Linear Regression

- Likelihood function of parameter w in training set D

$$p(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{w},\sigma) = \prod_{n=1}^{N} p(y^{(n)}|\boldsymbol{x}^{(n)};\boldsymbol{w},\sigma) = \prod_{n=1}^{N} \mathcal{N}(y^{(n)};\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}^{(n)},\sigma^2)$$

- MLE finds a set of parameters w such that the likelihood function $p(y|X;w,\sigma)$

$$\frac{\partial \log p(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{w},\sigma)}{\partial \boldsymbol{w}} = 0 \longrightarrow \boldsymbol{w}^{ML} = (\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}})^{-1}\boldsymbol{X}\boldsymbol{y}$$

Bayesian Perspective

北师香港浸会大学
BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY

- Maximum Likelihood:
  - When the training data is relatively small, overfitting will occur, resulting in inaccurate parameter estimation
  - Add prior (先验) knowledge to the parameters
- Bayesian Learning:
  - Consider the parameter w as a random variable
  - Objective: Given a set of observation data X, find the distribution $p(w|X)$ of parameter w
  - $p(w|X)$ is also called posterior distribution (后验分布)

Basic Idea of Probability

- Bayesian rule
  - The relationship between $p(y|x)$ and $p(x|y)$: $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$
  - Posterior, Likelihood, and prior

$$p(w|x) \propto p(x|w)p(w)$$

Posterior $\propto$ Likelihood$\times$ Prior

Maximum A Posteriori Estimation (MAP)

- Bayesian estimation is an interval estimation of parameters, that is, the distribution of parameters in an interval.
  - If you want to get an optimal parameter value, you can use the maximum a posteriori estimation (MAP).

$$\boldsymbol{w}^{MAP} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}; \sigma) p(\boldsymbol{w}; \nu)$$

$$\boldsymbol{w}^{MAP} = -\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}^{\mathsf{T}}\boldsymbol{w}\|^2 - \frac{1}{2\nu^2}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w}$$

正则化系数 $\lambda = \sigma^2/\nu^2$

2.6 Bias-Variance Decomposition

How to select a suitable model?

北师香港浸会大学
BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY

- Model Selection
  - Models with strong fitting ability are generally more complex and easy to overfitting.
  - If the complexity of the model is limited and the fitting ability is reduced, it may be underfitted.
- How to select a model?
  - The more complex the model, the lower the training error.
  - Model **cannot be selected according to the lowest training error**.
  - When **selecting a model, the test set is not visible**.

Bias-Variance Decomposition

北师香港浸会大学
BEIJING NORMAL-HONG KONG BAPTIST UNIVERSITY

- Expected risk:

$$\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim p_r(x,y)}\left[\left(y - f(x)\right)^2\right]$$

- The best model that ML algorithm can learn:

$$f^*(x) = \mathbb{E}_{y\sim p_r(y|x)}[y]$$

- Expected risk can decomposed into

$$\mathcal{R}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim p_r(\boldsymbol{x},y)}\left[\left(y - f^*(\boldsymbol{x}) + f^*(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2\right]$$

$$= \mathbb{E}_{\boldsymbol{x}\sim p_r(\boldsymbol{x})}\left[\left(f(\boldsymbol{x}) - f^*(\boldsymbol{x})\right)^2\right] + \epsilon$$

$$\epsilon = \mathbb{E}_{(\boldsymbol{x},y)\sim p_r(\boldsymbol{x},y)}\left[\left(y - f^*(\boldsymbol{x})\right)^2\right]$$

通常是由于样本分布以及噪声引起的，无法通过优化模型来减少。

真实目标：当前模型与最优模型之间的差距

Bias-Variance Decomposition

- Expected error

$$\mathcal{R}(f) = \mathbb{E}_{\boldsymbol{x}\sim p_r(\boldsymbol{x})}\Big[\mathbb{E}_{\mathcal{D}}\big[\big(f_{\mathcal{D}}(\boldsymbol{x}) - f^*(\boldsymbol{x})\big)^2\big]\Big] + \epsilon,$$

$$= (\text{bias})^2 + \text{variance} + \epsilon.$$

- Expected error is decomposed into

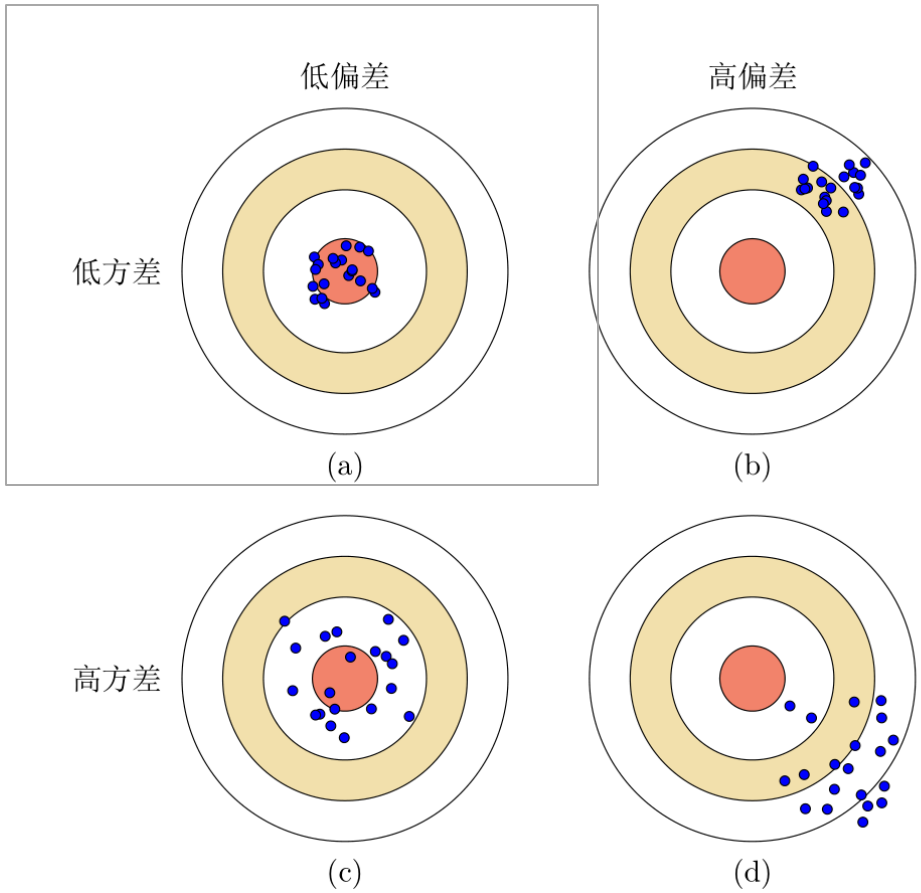
$$\mathcal{R}(f) = (\text{bias})^2 + \text{variance} + \varepsilon.$$

$$\mathbb{E}_{\mathbf{x}}\Big[\big(\mathbb{E}_{\mathcal{D}}\big[f_{\mathcal{D}}(\mathbf{x})\big] - f^*(\mathbf{x})\big)^2\Big]$$

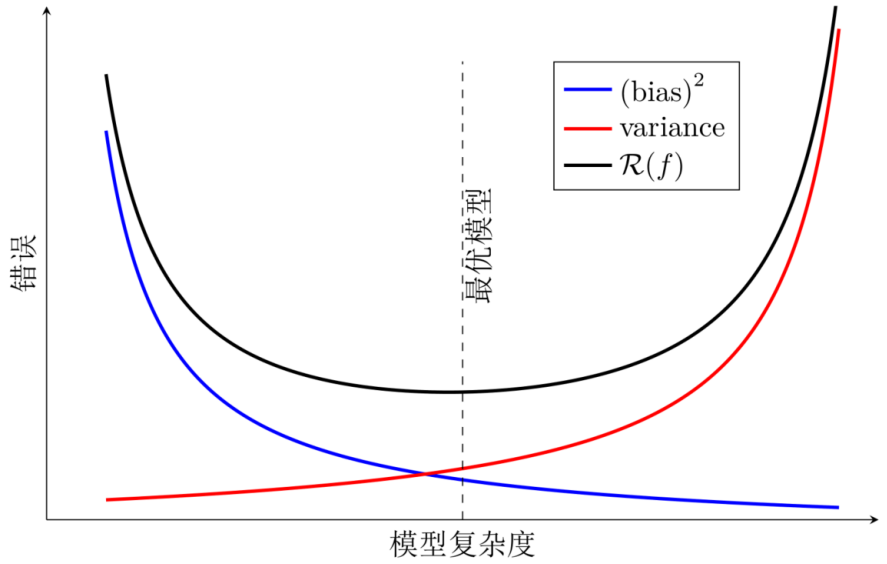
$$\mathbb{E}_{\mathbf{x}}\Big[\mathbb{E}_{\mathcal{D}}\Big[\big(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}\big[f_{\mathcal{D}}(\mathbf{x})\big]\big)^2\Big]\Big]$$

$$\mathbb{E}_{(\mathbf{x},y)\sim p_r(\mathbf{x},y)}\Big[\big(y - f^*(\mathbf{x})\big)^2\Big]$$

Model Selection: Deviation and Variance



Minimizing expected risk is equivalent to minimizing the sum of deviation and variance.



Ensemble model: an effective method to reduce variance