CS182 Introduction to Machine Learning

Recitation 5

2025.3.26

Outline

- review: probability & statistics
- Logistic Regression
- Metrics for binary classification

Review(Preview): Probability & Statistics

- 什么用到概率论与数理统计：
  - 概率论为机器学习提供了问题的假设
  - 回归和分类问题都可以描述为一个估计问题
  - 数据的分布往往服从正态分布
- 用哪些知识：
  - 常用的概率公式（条件概率、全概率、贝叶斯）
  - 常用的分布和他们的特殊性质（正态、泊松、两点、二项、均匀）
  - 常用的统计量（均值、方差、协方差）和他们的无偏估计

Probability

- $$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$
- Bayes' Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(A)$: prior probability of A
$P(B)$: Evidence
$P(A|B)$: posterior probability of A given B
$P(B|A)$: likelyhood (info of data given prior congnition)

Probability

- LOTP

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

> $\{B_1, B_2, \cdots, B_n\}$ is a partition of the sample space. 交集为空集, 并集为全集

Conditional Probability

> 条件概率也是概率

- Bayes' Rule

$$P(A|B,C) = \frac{P(B|A,C)P(A|C)}{P(B|C)}$$

- LOTP

$$P(A|C) = \sum_{i=1}^{n} P(A|B_i,C)P(B_i|C)$$

> Let $\hat{P}(\cdot) = P(\cdot|C)$

Independent & Exclusive Events 独立与互斥

- Independent:
  $P(A, B) = P(A)P(B)$
  $P(A|B) = P(A)$
  $P(B|A) = P(B)$
- Exclusive:
  $P(A \cup B) = P(A) + P(B)$
  i.e. $P(A \cap B) = 0$

Expectation

$$\mathbb{E}(X) = \int x f(x) dx$$

Linearity

- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- $\mathbb{E}(aX) = a\mathbb{E}(X)$
- $\nabla \mathbb{E}(f) = \mathbb{E}(\nabla f)$
- $X \perp\!\!\!\perp Y \Rightarrow \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$

  LOTUS(Law of the Unconscious Statistician)

- $\mathbb{E}(g(X)) = \int g(x) f(x) dx$

> 条件期望也是期望

Variance

- $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$
- $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

If X and Y are independent, then $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$
方差没有线性性

- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$
- $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$

Covariance

- $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$
- $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$

> $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ if X and Y are independent

Correlation 相关性

- $\rho = \mathrm{Corr}(X, Y) = \dfrac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \in [-1, 1]$

只能刻画**线性**相关性, 且正负相关程度相同(正负相关).
X,Y独立, 则$\rho_{X,Y} = 0$. 但是$\rho_{X,Y} = 0$**不一定**独立.

> e.g. $Y = X^2, X \sim N(0,1) \Rightarrow \mathbb{E}(X) = 0, \mathrm{Var}(X) = 1, \mathbb{E}(Y) = \mathbb{E}(X^2) = 1$
> $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$

$Gaussian$分布独立 $\Leftrightarrow$ 不相关.

Iterated Expectation(Adam's law)

For any r.v. X, Y and Z:

$$\mathbb{E}_X\left[\mathbb{E}_{Y|X}(Y|X)\right] = \mathbb{E}_Y(Y)$$

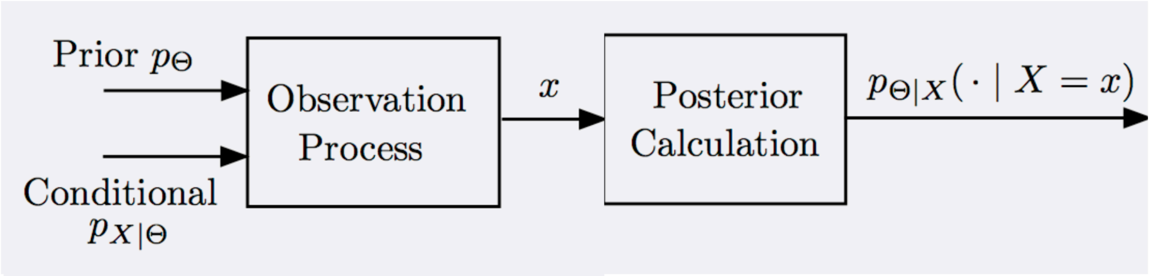
$$\mathbb{E}_X\left[\mathbb{E}_{Y|X}(Y|X,Z)|Z\right] = \mathbb{E}_Y(Y|Z)$$

Law of Total Variance(Eve's law)

$$\mathrm{Var}(Y) = \mathbb{E}_X\left[\mathrm{Var}(Y|X)\right] + \mathrm{Var}(\mathbb{E}_{Y|X}(Y|X))$$

MAP: maximum a posteriori 最大后验分布

Bayesian 统计学派



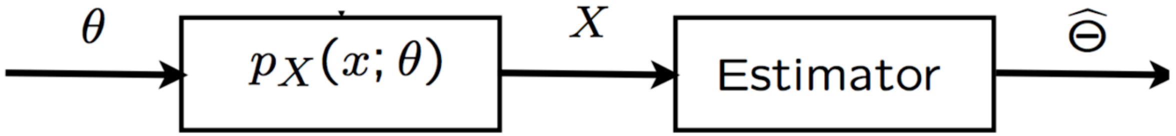
贝叶斯学派把一切变量看作随机变量, 利用过去的知识和抽样数据,将概率解释为信念度(degree of belief), 不需要大量的实验

Θ是个随机变量, 根据样本的具体情况来估计参数, 使样本发生发可能性最大(与时俱进,不断更新)

$$\hat{\theta} = \arg\max_{\theta} p(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P_{\Theta}(\theta)$$

MLE: maximum a likelihood 最大似然估计

Frequentist 统计学派



频率学派把未知参数看作普通变量(固定值), 把样本看作随机变量, 仅仅利用抽样数据, 频率论方法通过大量独立实验将概率解释为统计均值(大数定律)

$$\hat{\theta} = \arg\max_{\theta} p(\mathcal{D};\theta) \text{ or } \hat{\theta} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

$$\overset{i.i.d.}{=} \arg\max_{\theta} \prod_{i=1}^{n} p(y_i|\theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log p(y_i|\theta)$$

Conjugate Prior 共轭先验

Sample Space	Sampling Dist.	Conjugate Prior	Posterior
$\mathcal{X} = \{0, 1\}$	Bernoulli(θ)	Beta(α, β)	Beta$(\alpha + n\overline{X}, \beta + n(1 - \overline{X}))$
$\mathcal{X} = \mathbb{Z}_+$	Poisson(λ)	Gamma(α, β)	Gamma$(\alpha + n\overline{X}, \beta + n)$
$\mathcal{X} = \mathbb{Z}_{++}$	Geometric(θ)	Gamma(α, β)	Gamma$(\alpha + n, \beta + n\overline{X})$
$\mathcal{X} = \mathbb{H}_K$	Multinomial(θ)	Dirichlet(α)	Dirichlet$(\alpha + n\overline{X})$

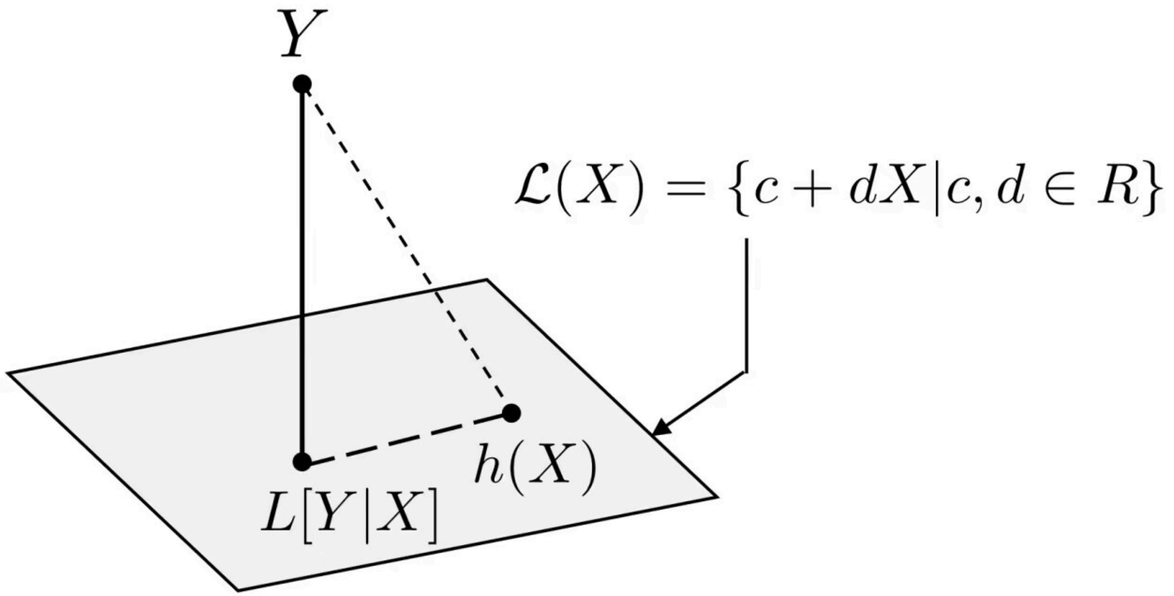
Conjugate Prior 共轭先验

Sampling Dist.	Conjugate Prior	Posterior
Uniform(θ)	Pareto(ν_0, k)	Pareto $\left(\max\{\nu_0, X_{(n)}\}, n+k\right)$
Exponential(θ)	Gamma(α, β)	Gamma$(\alpha + n, \beta + n\overline{X})$
$N(\mu, \sigma^2)$, known σ^2	$N(\mu_0, \sigma_0^2)$	$N\left(\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{X}}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right)$
$N(\mu, \sigma^2)$, known μ	InvGamma(α, β)	InvGamma $\left(\alpha + \frac{n}{2}, \beta + \frac{n}{2}\overline{(X-\mu)^2}\right)$
$N(\mu, \sigma^2)$, known μ	ScaledInv-$\chi^2(\nu_0, \sigma_0^2)$	ScaledInv-χ^2 $\left(\nu_0 + n, \frac{\nu_0\sigma_0^2}{\nu_0 + n} + \frac{n\overline{(X-\mu)^2}}{\nu_0 + n}\right)$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, known $\boldsymbol{\Sigma}$	$N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$	$N\left(\mathbf{K}\left(\boldsymbol{\Sigma}_0^{-1}\mu_0 + n\boldsymbol{\Sigma}^{-1}\overline{X}\right), \mathbf{K}\right)$, $\mathbf{K} = \left(\boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, known $\boldsymbol{\mu}$	InvWishart$(\nu_0, \mathbf{S}_0)$	InvWishart$(\nu_0 + n, \mathbf{S}_0 + n\overline{\mathbf{S}})$, $\overline{\mathbf{S}}$ sample covariance

Linear Least Square Estimate (LLSE)

拟合函数为线性: $h(X) = a + bX$, 则 minimizes $\mathbb{E}\left[(Y - a - bX)^2\right]$ 的函数为

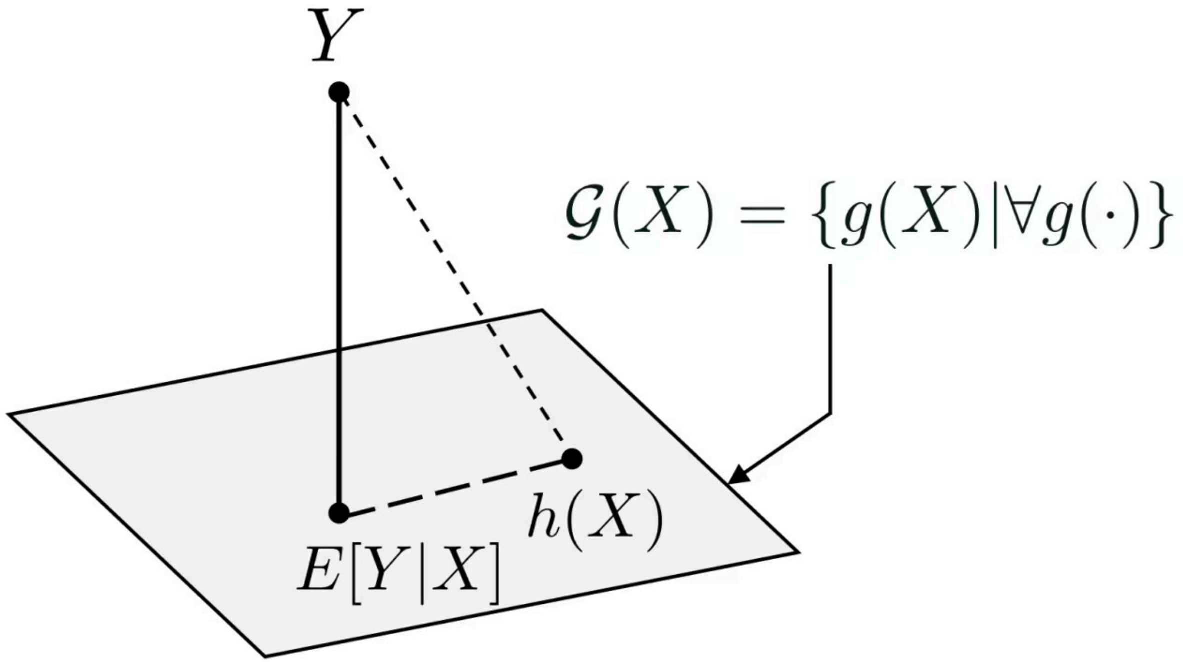
$$L[Y|X] = \mathbb{E}(Y) + \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}(X - \mathbb{E}(X))$$



Minimum Mean Square Error (MMSE)

拟合函数为**任意函数**h(X), 则 minimizes $\mathbb{E}\left[(Y-h(X))^2\right]$的函数为

$$h(X)=\mathbb{E}(Y|X)$$



Concentration Inequalities

For any r.v. X, and any constant $a > 0$:

- Markov Inequality

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

- Chebyshev Inequality

$$P(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$$

- Chernoff's Inequality

$$P(X \geq a) \leq \frac{\mathbb{E}(e^{tX})}{e^{ta}}, \forall t > 0$$

Probability Inequalities

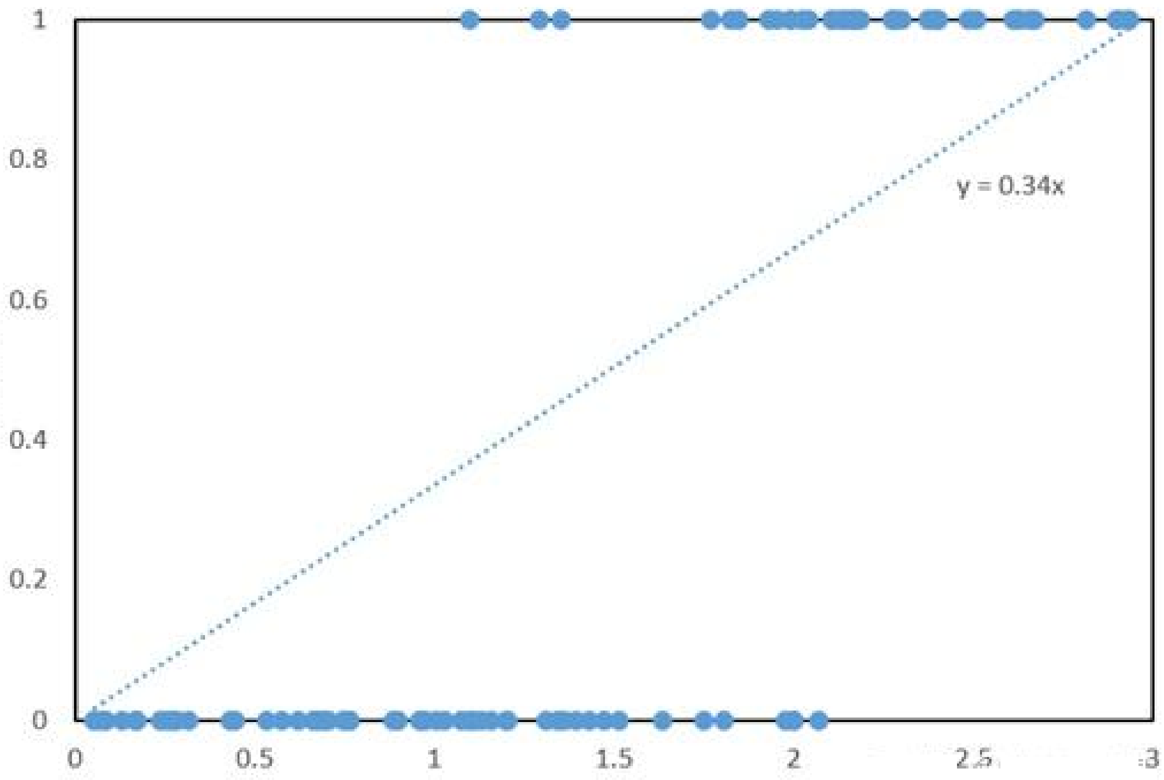
- Hoffding Lemma: $\mathbb{E}(X) = 0, a \leq X \leq b, \forall \lambda > 0$

$$\mathbb{E}(e^{\lambda X}) \leq e^{\frac{1}{8}\lambda^2(b-a)^2}$$

- Hoffding Bound: $X_1, \ldots, X_n$ independent, $\mathbb{E}(X_i) = \mu, a \leq X_i \leq b, \forall \epsilon > 0$

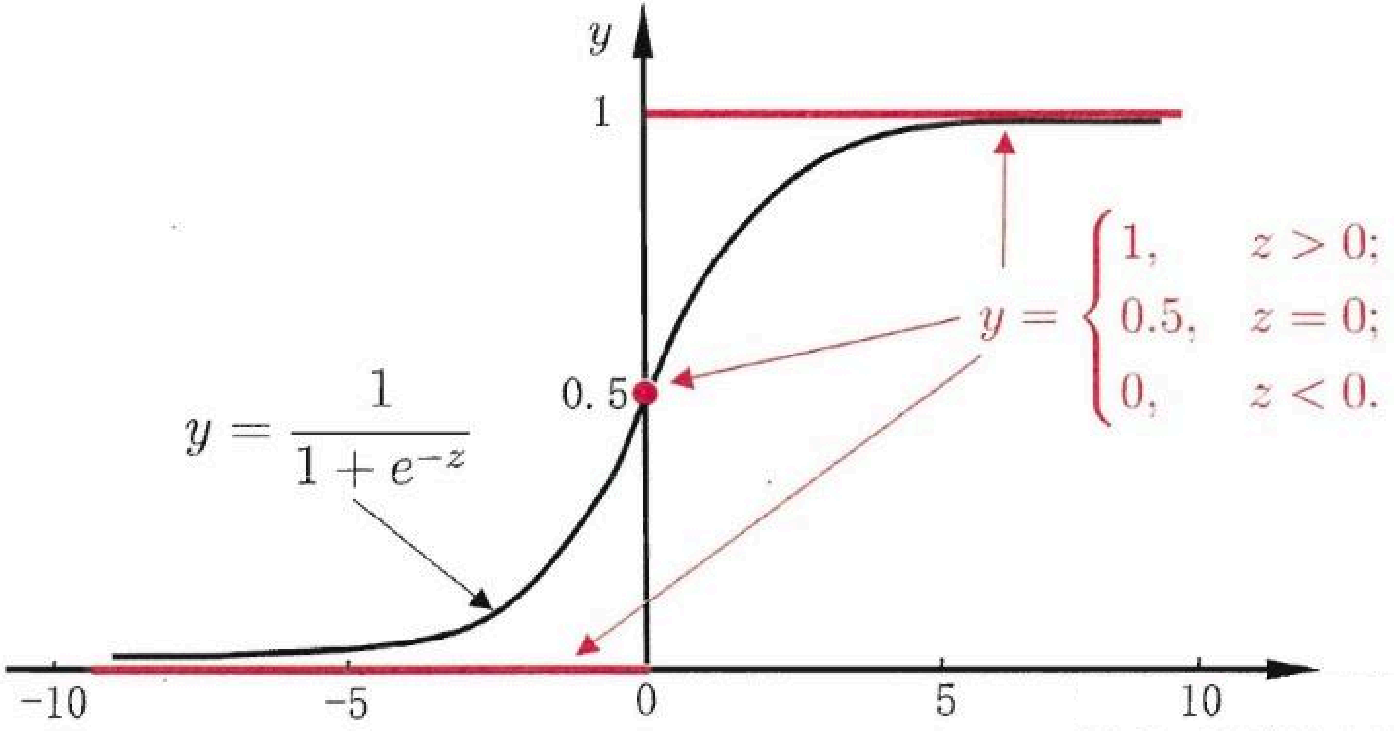
$$P(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq \epsilon) \leq 2\exp\left(-\frac{2n\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Logistic Regression



https://zhuanlan.zhihu.com/p/139122386

Sigmoid function



$$p(y_i = 1|x_i;\theta) = \sigma(\theta^\top x_i) = \frac{1}{1+e^{-\theta^\top x_i}}$$

$$p(y_i = 0|x_i;\theta) = 1 - \sigma(\theta^\top x_i) = \frac{e^{-\theta^\top x_i}}{1+e^{-\theta^\top x_i}}$$

Logistic Regression

$$p(y_i = 1|x_i;\theta) = \sigma(\theta^\top x_i) = \frac{1}{1+e^{-\theta^\top x_i}}$$

$$p(y_i = 0|x_i;\theta) = 1 - \sigma(\theta^\top x_i) = \frac{e^{-\theta^\top x_i}}{1+e^{-\theta^\top x_i}}$$

$$\Rightarrow p(y_i|x_i;\theta) = \left[\sigma(\theta^\top x_i)\right]^{y_i} \cdot \left[1 - \sigma(\theta^\top x_i)\right]^{1-y_i}$$

$$\begin{aligned}
\hat{\theta} &= \arg\max_\theta \prod_{i=1}^{n} p(y_i|x_i;\theta) \\
&= \arg\max_\theta \prod_{i=1}^{n} \left[\sigma(\theta^\top x_i)\right]^{y_i} \cdot \left[1 - \sigma(\theta^\top x_i)\right]^{1-y_i} \\
&= \arg\max_\theta \sum_{i=1}^{n} y_i \log\left[\sigma(\theta^\top x_i)\right] + (1-y_i)\log\left[1 - \sigma(\theta^\top x_i)\right]
\end{aligned}$$

Multi-class Logistic Regression

sigmoid -> softmax

$$p(y_i = k|x_i; W) = \frac{e^{w_k^\top x_i}}{\sum_{j=1}^{K} e^{w_j^\top x_i}}$$

$$\mathcal{L}(W) = \prod_{i=1}^{n} \prod_{k=1}^{K} p(y_i = k|x_i; W)^{y_{ik}}$$

Or commonly as

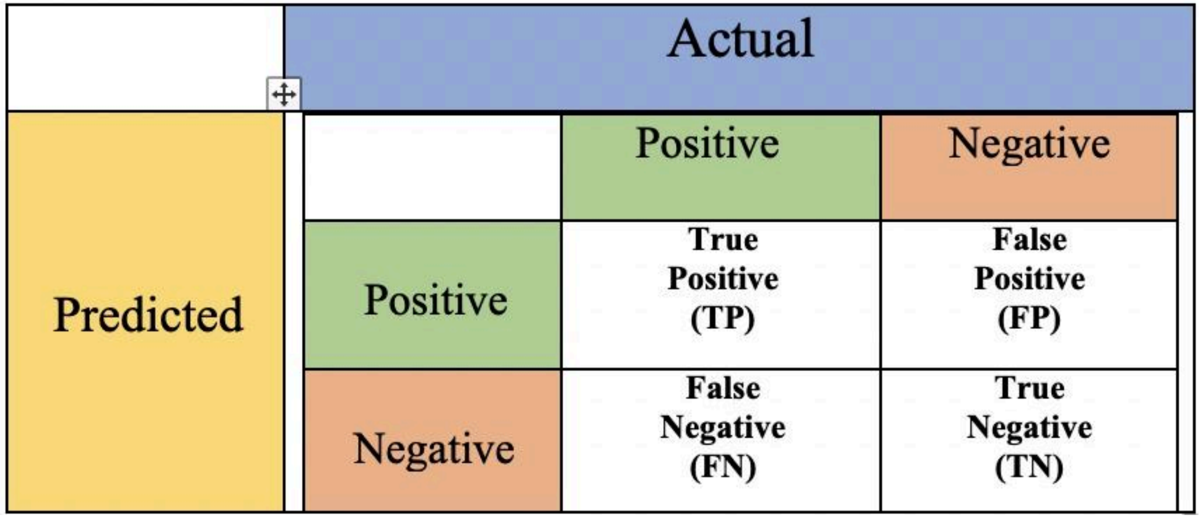
$$\mathcal{L}(W) = \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \log p(y_i = k|x_i; W)$$

其中$y_{ik} = 1$表示第i个样本的真实标签是第k类, 否则 $y_{ik} = 0$

Metrics for binary classification

- Accuracy
- Precision
- Recall
- F1 Score
- ROC Curve
- ROC-AUC

Metrics for binary classification



		Actual	
		Positive	Negative
Predicted	Positive	**True Positive (TP)**	**False Positive (FP)**
	Negative	**False Negative (FN)**	**True Negative (TN)**

True Positive(TP): 真阳性

False Positive(FP): 假阳性(错误的判断为正样本)

True Negative(TN): 真阴性

False Negative(FN): 假阴性(错误的判断为负样本)

> reference: https://zhuanlan.zhihu.com/p/364253497

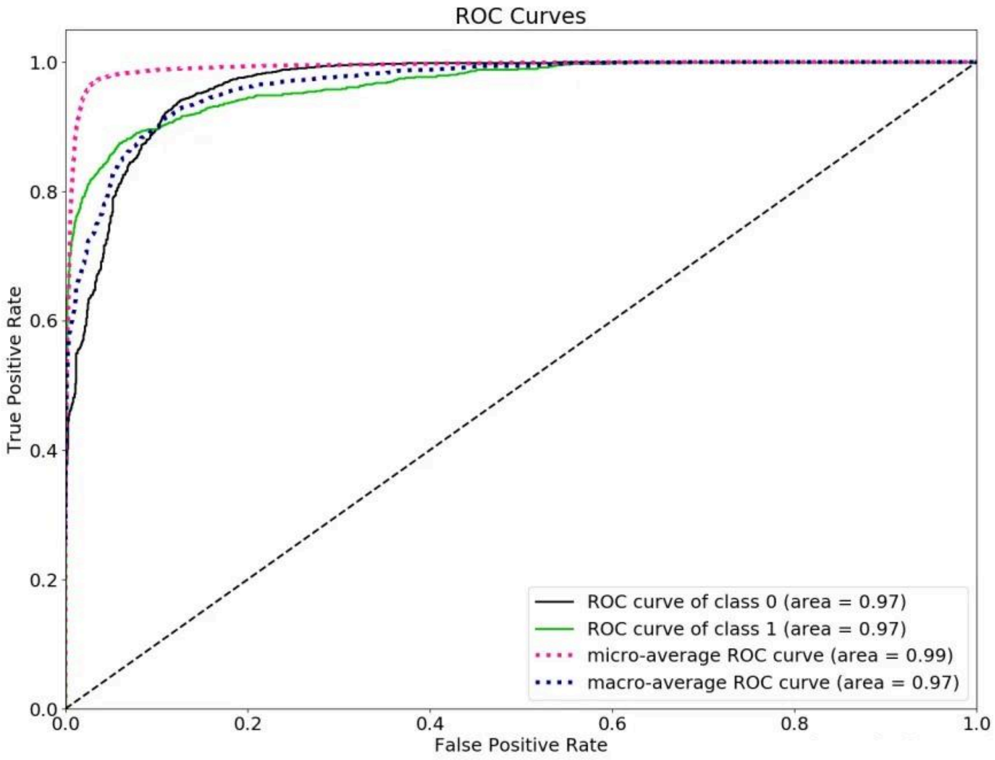
Metrics for binary classification

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

- Accuracy: 正确率 = $\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$
- Precision: 精准度 = $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$
- Recall: 召回率 = $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
- F1 Score = $\frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$

Metrics for binary classification

- ROC (Receiver Operating Characteristic)



ROC-AUC(area-under-curve)