Pattern Recognition and Machine Intelligence: Assignment 1

Due on **Feb. 28, 2024** at 23:59 pm

Dr. Xiaojuan Qi

TA: Runyu Ding Haozheng Wan

Problem 1

[64 pts] Please select all of the choices that apply. Please note that there may be more than one correct answer.

B (a) **[8 pts]** What is the difference between the generative model and the discriminative model?

A. A generative model directly learns the probability of the output labels given the input data, while a discriminative model learns the conditional probability density of the input data given the label.

B. A generative model models data distribution within a class and can make predictions with the Bayes rule, while a discriminative model directly learns the decision boundary (posteriori) between different classes.

C. A generative model learns the decision boundary between different classes, while a discriminative model models how the data is generated and can make predictions with the Bayes rule.

ABC (b) **[8 pts]** Which descriptions about supervised learning and unsupervised learning are correct?

A. Supervised learning trains on labeled data, meaning that each data point in the training set has an associated human annotated label.

B. A unsupervised learning model is trained on unlabeled data, meaning that the training set does not include any human labels.

C. To collect data for training a supervised learning model, we also need to collect the corresponding labels of the data samples.

AC (c) **[8 pts]** What are the pros and cons of using multiple features instead of single features?

A. Using multiple features allows the model to capture more complex relationships between the input and output variables, which might lead to better predictive performance on new data.

B. Using multiple features can reduce the computational cost of training the model.

C. As the number of features increases, the number of possible combinations of features grows exponentially, making it harder to find a good model and leading to overfitting in some cases if the amount of data is not sufficient.

CD (d) **[8 pts]** What is true about dealing with overfitting and model generalization in machine learning?

A. The objective is to achieve the highest accuracy on the training data.

B. A more complex model is always preferable.

C. Striking a balance between fitting the training data well and generalizing to unseen test data is important.

D. More training data can help avoid overfitting.

BC (e) **[8 pts]** Which of following descriptions about MLE and MAP are correct?

A. MLE always provides the best estimate of model parameters.

B. MLE finds the value of the parameters that maximizes the likelihood of the training data, while MAP finds the value of the parameter that maximizes the posterior distribution of the parameters.

C. The posterior distribution takes into account both the likelihood function and a prior distribution over the parameter, which represents our prior belief about the value of the parameter.

Problem 1 continued on next page...

A (f) **[8 pts]** Given a dataset of coin toss outcomes, where 'H' represents heads, and 'T' represents tails. The dataset consists of 100 tosses, and you observed 62 heads (H) and 38 tails (T). Which value of P(H) should you estimate using Maximum Likelihood Estimation (MLE) for this dataset?

A. P(H) = 0.62 ✓

B. P(H) = 0.38

C. P(H) = 0.5

看清楚题！

NB 假设：$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ 在给定 w 时，$x_1/x_2/\cdots/x_n$ 间是条件独立的

$p(x|w) = \prod_{i=1}^{n} p(x_i|w)$

BC (g) **[8 pts]** Which of the following statements correctly describes the Bayesian classifier and the naive Bayesian classifier?

A. The Bayesian classifier is a simplified version of the Naive Bayes classifier. ✗

B. The Naive Bayes classifier assumes that different dimensions of the input feature are conditionally independent of each other given the class. ✓

C. The Naive Bayes classifier is more computationally efficient than the Bayesian classifier. ✓

AB (h) **[8 pts]** Which of the following statements correctly describes the process of Bayesian classifier?

A. In the training phase, we need to estimate the prior and likelihood of each class (i.e., the class-conditional probability density of each class) from the training data. ✓

B. In the testing phase, we use the Bayes rule to obtain the posterior according to prior and likelihood obtained from the training phase. ✓

C. We assign a sample to the class that has the largest likelihood. ✗

Problem 2

$N(\mu, \sigma^2)$

[36 pts] Suppose you have a dataset of samples $D = \{x_1, x_2, \ldots, x_n\}$ from a Gaussian distribution with unknown mean μ and known variance σ^2. The samples are independently and identically distributed (i.e. iid.). You want to estimate the value of μ from the set of samples.

naïve assumption

(a) **[10 pts]** Write down the likelihood function for this problem, i.e. $p(D \mid \mu)$.

(b) **[14 pts]** Use Maximum Likelihood Estimation (MLE) to find the estimate for μ.

(c) **[12 pts]** You have prior knowledge that the value of μ follows a gaussian distribution with a mean of 80 and a variance of 16. Use Maximum a Posterior (MAP) to find the estimate for μ.

$$p(\mu) = \frac{1}{4\sqrt{2\pi}} \exp[-\frac{(\mu - 80)^2}{32}]. \tag{1}$$

Hint: Find μ that maximizes the posterior of μ. You may consider the the posterior of μ given the data D: $p(D \mid \mu)p(\mu)$, and apply a log function.

Solution

a) $P(D|\mu) = \prod_{i=1}^{n} P(x_i|\mu)$

$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}\cdot(\frac{x_i-\mu}{\sigma})^2}$

b) MLE - for μ

A: $\mu^* = \arg\max_{\mu} P(D|\mu, \sigma^2)$

$= \arg\max_{\mu} \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \sum_{i=1}^{n} -\frac{1}{2}\cdot(\frac{x_i-\mu}{\sigma})^2$

$= \arg\max_{\mu} -\frac{1}{2}\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2}$

$= \arg\max_{\mu} -\frac{1}{2}\sum_{i=1}^{n} (x_i-\mu)^2$

$= \arg\max_{\mu} F(x, \mu)$

let. $\frac{\partial F(x,\mu)}{\partial \mu} = \sum_{i=1}^{n} (-\frac{1}{2})\cdot(2\mu - 2x_i) = \sum_{i=1}^{n} (x_i - \mu) = 0$

$\therefore \mu^* = \frac{\sum_{i=1}^{n} x_i}{n}$

c) $p(\mu) = \frac{1}{4\sqrt{2\pi}} \cdot e^{-\frac{(\mu-80)^2}{32}}$, MAP

A: $\mu^* = \arg\max_{\mu} p(\mu|D)$

$= \arg\max_{\mu} p(D|\mu) \cdot p(\mu) = \arg\max_{\mu} \ln p(D|\mu) + \ln p(\mu)$

$= \arg\max_{\mu} \sum_{i=1}^{n} \ln p(x_i|\mu) + \ln p(\mu)$

$= \arg\max_{\mu} -\frac{1}{2} \cdot \sum_{i=1}^{n} (x_i-\mu)^2 - \frac{(\mu-80)^2}{32} = \arg\max_{\mu} F(x,\mu)$

let: $\frac{\partial F(x,\mu)}{\partial \mu} = \sum_{i=1}^{n} \left(-\frac{1}{2}\right) \cdot (2\mu - 2x_i) - \frac{1}{32} \cdot (2\mu - 160) = 0$

$\therefore \mu^* = \frac{5 + \sum_{i=1}^{n} x_i}{n + \frac{1}{16}}$

$p(x_i|\mu) = \frac{1}{\sqrt{2\pi}\,6} \cdot e^{-\frac{1}{2}\left(\frac{x_i-\mu}{6}\right)^2}$

$-\frac{1}{2} \cdot \left(\frac{x_i-\mu}{6}\right)^2$

化简错误！不可以跳步!! 一步步来哦！

$\arg\max_{\mu} (a+b)$ 只有 a,b 中和 μ 无关的可以去掉，系数必须保留!!!

补充学习

过拟合

原因

过拟合是机器学习中常见的问题，其原因可以归结为以下几点：

1. 数据不足：当训练数据较少时，模型容易过度拟合训练集中的噪声和异常值，而无法很好地泛化到新的数据上。

2. 模型复杂度过高：当模型的复杂度过高时，它具备更多的参数和更强的拟合能力，容易记住训练数据的细节，但也更容易过拟合。例如，高阶多项式模型可以很好地拟合训练数据，但在新数据上的表现可能较差。

3. 特征选择不当：如果选择的特征不具有很好的表示能力，或者包含大量无关的特征，模型容易过拟合训练数据，无法很好地泛化到新的数据上。

4. 训练集和测试集分布不一致：当训练集和测试集的数据分布不一致时，模型可能会过拟合训练集的特定模式，而在测试集上表现较差。

5. 数据噪声：训练数据中的噪声和异常值可以导致模型过度拟合这些不真实的数据，从而降低模型的泛化能力。

了解这些原因有助于我们在建模过程中更好地理解和处理过拟合问题，并采取相应的措施来减轻其影响。

18/03/2024, 15:19:47

解决

有很多可以减轻过拟合的方法，其中一些常用的方法如下：

1. 增加训练数据：增加训练数据可以减轻过拟合的问题，因为更多的数据可以提供更多的信息和多样性，使得模型更容易地捕捉数据的整体特征和规律。

2. 正则化：正则化是一种通过对模型的参数进行惩罚来避免过拟合的方法。常见的正则化方法包括L1正则化和L2正则化。

3. 交叉验证：交叉验证是一种将数据集划分为多个子集的方法，其中一个子集被用作验证集，其余子集被用作训练集。通过多次交叉验证来评估模型的性能，可以减少模型对特定数据集的过拟合。

4. 剪枝：剪枝是一种通过删除模型中的一些决策树节点来减少模型复杂度的方法，从而减轻过拟合的问题。剪枝可以应用于决策树、神经网络等模型。

5. 集成学习：集成学习是一种将多个模型组合在一起来提高性能的方法。常见的集成学习方法包括随机森林、Boosting等。

这些方法通常都可以帮助减轻过拟合的问题，但并非一定适用于所有情况，需要根据具体情况选择合适的方法。

18/03/2024, 15:18:16

tip：数据不足 & 训/测分布不一致
会同时 overfit / underfit !!!

欠拟合

原因

欠拟合是机器学习中另一个常见的问题，其原因可以归结为以下几点：

1. 模型复杂度过低：当模型的复杂度过低时，它可能无法很好地拟合训练数据的复杂模式和关系。例如，线性模型可能无法捕捉到非线性关系。

2. 特征选择不足：如果选择的特征不具备很好的表示能力，或者忽略了重要的特征，模型可能无法充分利用数据的信息，导致欠拟合。

3. 数据噪声和异常值：当训练数据中存在大量噪声或异常值时，模型可能会受到干扰，无法很好地拟合数据的真实模式。

4. 训练集和测试集分布不一致：如果训练集和测试集的数据分布不一致，模型可能无法很好地泛化到新的数据上，导致欠拟合。

5. 数据量不足：当训练数据较少时，模型可能无法学习到数据的整体特征和规律，导致欠拟合。

欠拟合的问题表明模型无法很好地拟合训练数据，通常可以通过增加模型复杂度、选择更好的特征、增加训练数据量等方法来改善欠拟合问题。

解决

解决欠拟合的方法有以下几种：

1. 增加模型复杂度：当模型无法拟合训练数据时，可以考虑增加模型的复杂度，例如增加神经网络的层数或节点数、增加多项式的阶数等。

2. 特征工程：通过选择更好的特征、提取更多的特征或者对特征进行组合，可以提高模型的拟合能力，从而减轻欠拟合的问题。

3. 增加训练数据量：增加训练数据量可以提高模型的泛化能力，从而减轻欠拟合的问题。

4. 减小正则化系数：如果模型采用了正则化方法，可以尝试减小正则化系数，以降低对模型的惩罚，提高模型的拟合能力。

5. 改变模型架构：可以尝试改变模型的架构，例如从线性模型转换为非线性模型，或者从单个模型转换为集成学习模型。

6. 调整超参数：调整模型的超参数，例如学习率、批次大小等，可以提高模型的拟合能力。

需要根据具体情况选择合适的方法来解决欠拟合问题。如果欠拟合问题较为严重，可能需要采用多种方法组合使用，以达到更好的效果。