2980 反垃圾邮件

一、 处理步骤

1. 从 EML 文件解析出数据：邮件标题文本，发件日期，发件人邮箱地址，IP，收件人，抄送人
2. 数据预处理：
    - 去除重复标题文本数据
    - 纠错部分明显误分邮件，例如色情邮件分为正常，多益相关邮件分为垃圾邮件
    - 去除标题中无意义字符
    - 文本分词
    - 滤除出现频率较低的字符
    - 格式化接受日期
3. 邮件数据向量化
    - 日期向量化：
        - 将日期当作离散值处理，采用 one-hot 编码将月份、该月哪一天、该周哪一天等编码为二进制数据，例如每周的周三，可以编码为[0,0,1,0,0,0,0].
        - 假设日期是周期性连续特征值，相邻的两天是连续的 24 小时周期，相邻的两小时是连续的 60 分钟周期，那么我们可以算出每个周期值得正弦和余弦值作为特征值
    - 文本向量化
        - 词袋模型
        - TF-IDF
        - 词嵌入：Skip-gram / CBOW
4. 高维度处理&特征值选择：
    - 滤除方差较小的特征
    - PCA（主成份分析），挑选数据集中最大偏差的成分
    - 随机采样数据，平衡二分类样本比例
    - 对数值变量做相关性分析，滤除一些相关性较大的列
    - 增量训练
5. 机器学习模型：
    - 朴素贝叶斯
    - 逻辑回归
    - 神经网络
6. 评估指标

    测试准确率：

    测试样本中，预测标签与实际标签相同的概率值

    精准率：

针对预测标签为阳性（垃圾邮件）的样本，其中被正确预测为阳性的概率值

召回率：

针对实际标签为阳性的样本，其中被正确预测为阳性的概率值

F1：

精准率与召回率的调和平均值，由于精准度或者召回率无法单独代表模型的优劣。例如将所有样本全部预测为阳性，召回率即为 100%，但是模型是很差的。

混淆矩阵：

预测标签 / 真实标签	正常（Negative）	垃圾（Positive）
Negative	TN	FP
Positive	FN	TP

TN：正确预测为正常邮件的数量

FP：错误预测为垃圾邮件的数量

FN：错误预测为正常邮件的数量

TP：正确预测为垃圾邮件的数量

ROC 曲线：

一种显示分类模型在所有分类阈值下的效果图表，以真正例率 (召回率，TPR)·为 Y 轴，假正例率（FPR）为 X 轴，绘制采用不同分类阈值时的 TPR 与 FPR。降低分类阈值会导致将更多样本归为正类别，从而增加假正例和真正例的个数

TPR = TP / TP + FN

FPR = FP / FP + TN

AUC:

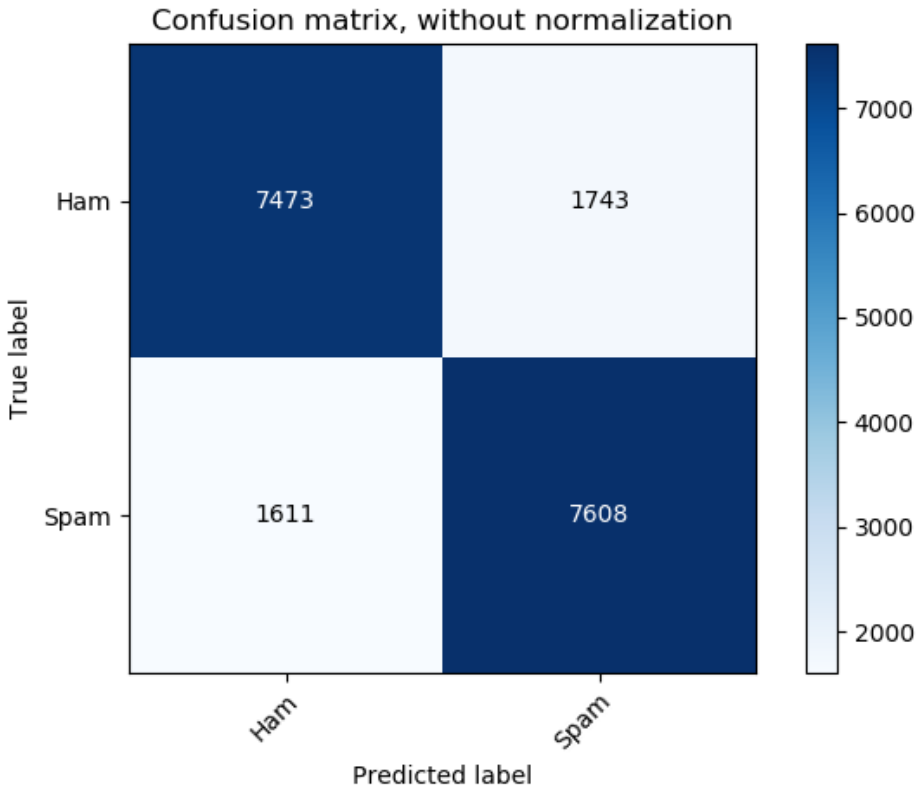
ROC 曲线下面积，它对所有可能的分类阈值的效果进行综合衡量，可看作模型将某个随机正类别样本排列在某个随机负类别样本之上的概率

二、 特征挖掘与模型训练

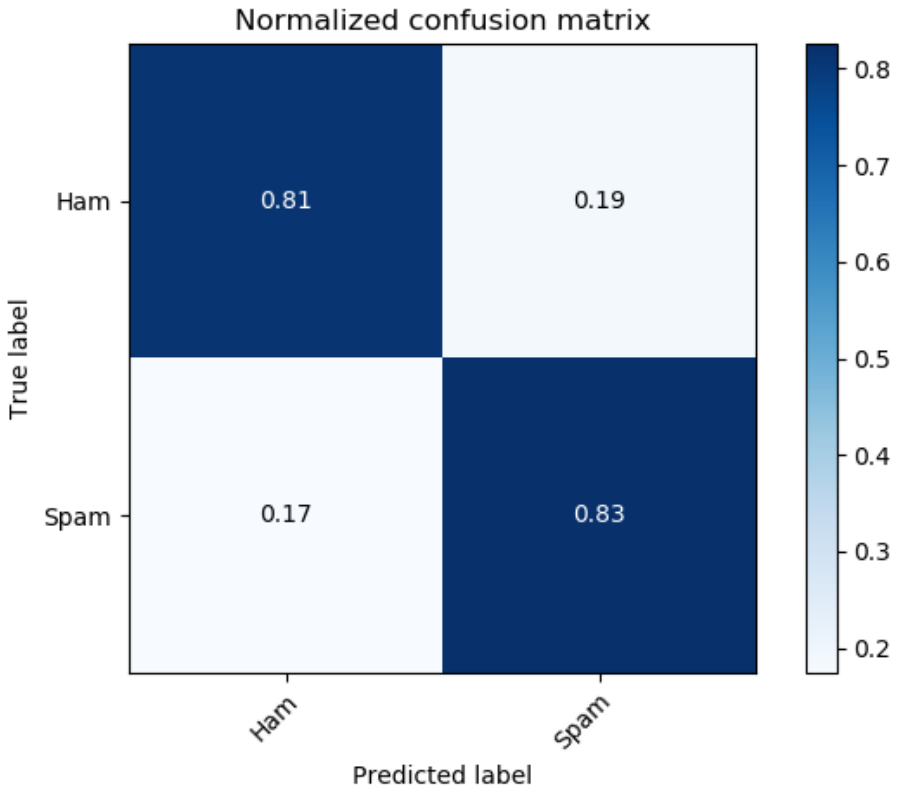
在这一章，我们比较了不同的邮件数据特征结合给模型指标带来的影响，以图表的形式展示，首先我们仅仅从邮件的文本标题中挖掘特征，然后逐步的将一些邮件中的其它数据做处理并添加到训练数据的特诊维度中去，例如：日期、IP 等可能帮助分类的信息。训练过程中，我们控制模型不变，但是 sklearn 内置的贝叶斯模型对数据有不同的要求，例如 Bernoulli Naïve Bayes 假设数据服从二项分布，Multinomial Naïve Bayes 假设数据是离散的，例如电影打分分布在 1-5 之间，Gaussian Naïve Bayes 假设数据服从正态分布。总的来说，模型为贝叶斯，但对数据分布的假设不同。

2.1 文本标题：

预处理：文本词袋模型
机学模型：Bernoulli Naïve Bayes
评估图表：

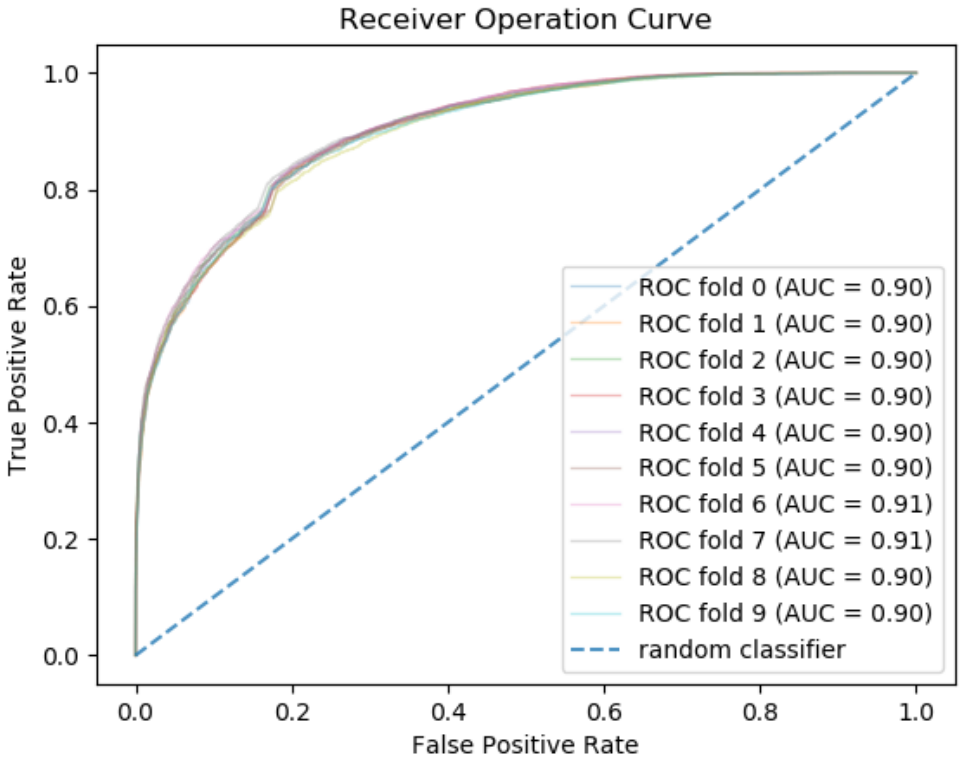


图表 2.1.1 混淆矩阵



图表 2.1.2 标准化后的混淆矩阵

Ham-正常邮件　　Spam-垃圾邮件



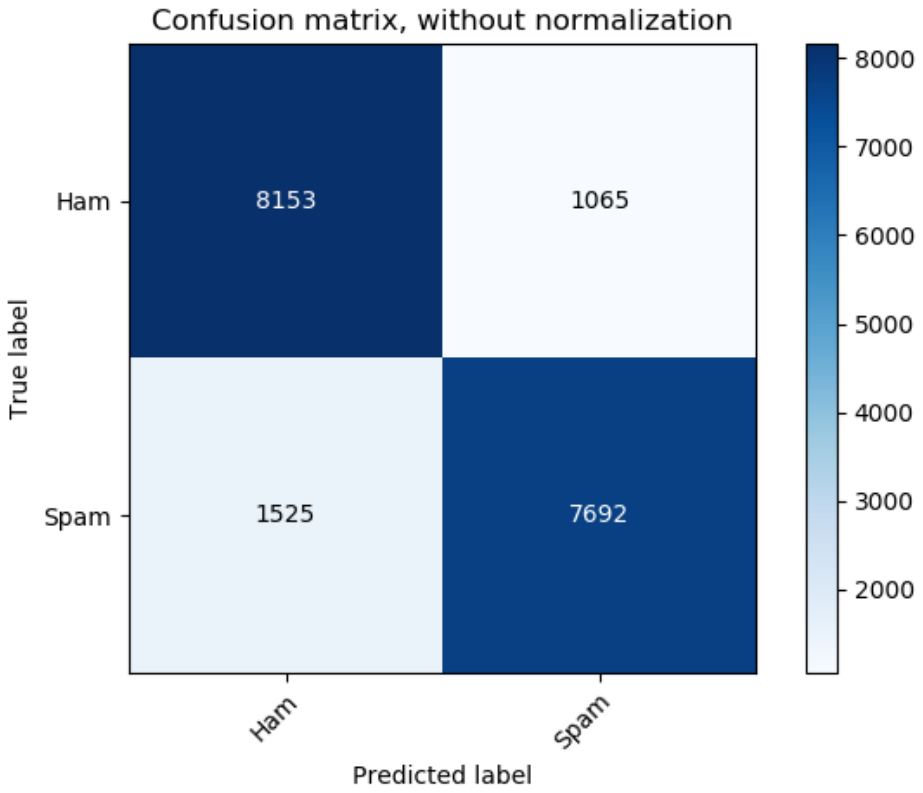
图表 2.1.3 10 折交叉验证 ROC 曲线

2.2　文本标题+时间（假设离散）

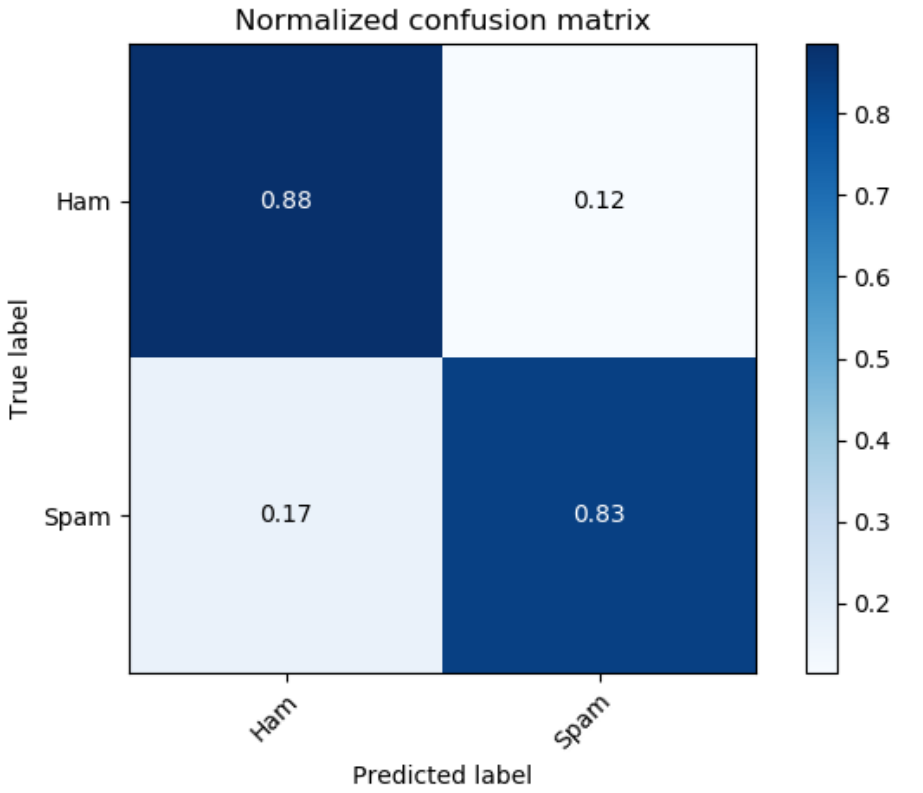
预处理：词袋模型处理文本标题，假设时间为离散变量，例如星期几，可理解为分布与 1-7 之间的离散值，编码为[0,0,0,0,0,1, 0](星期六)。

机学模型：Bernoulli Naïve Bayes

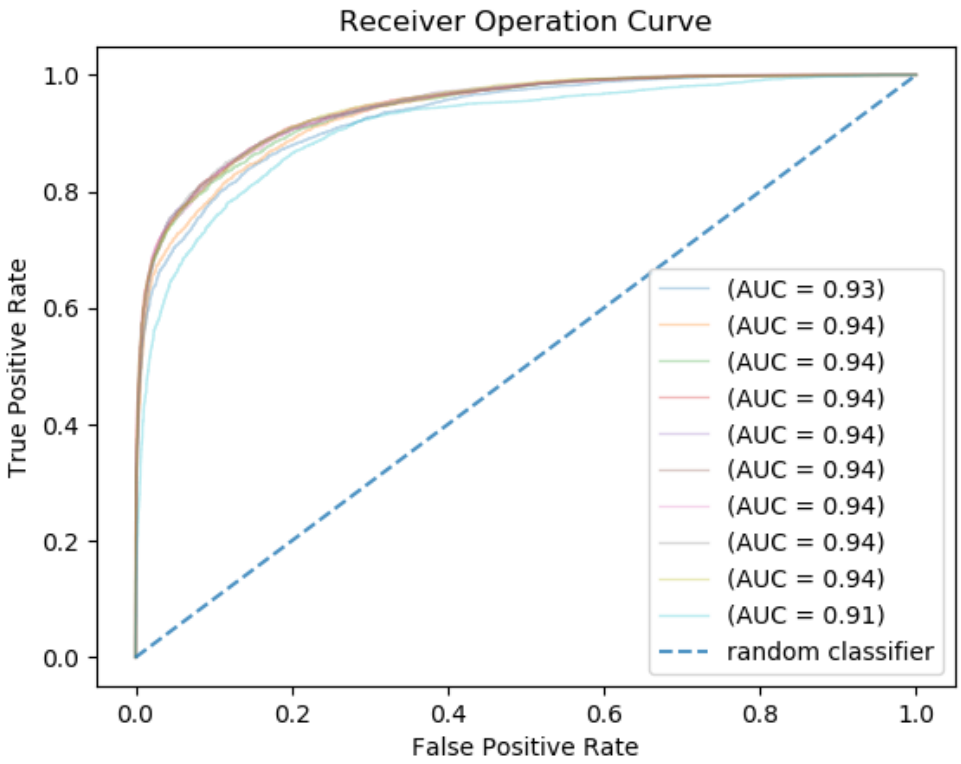
评估图表：



图表 2.2.1 混淆矩阵



图表 2.2.3 标准化混淆矩阵

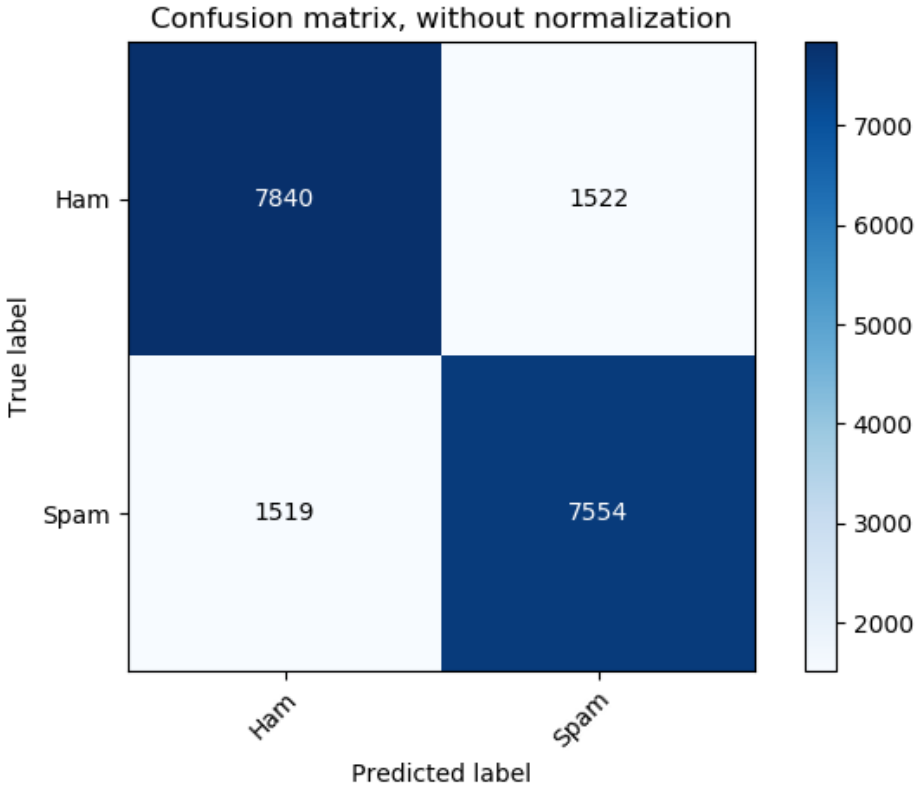


图表 2.2.3 10 折交叉验证 ROC 曲线

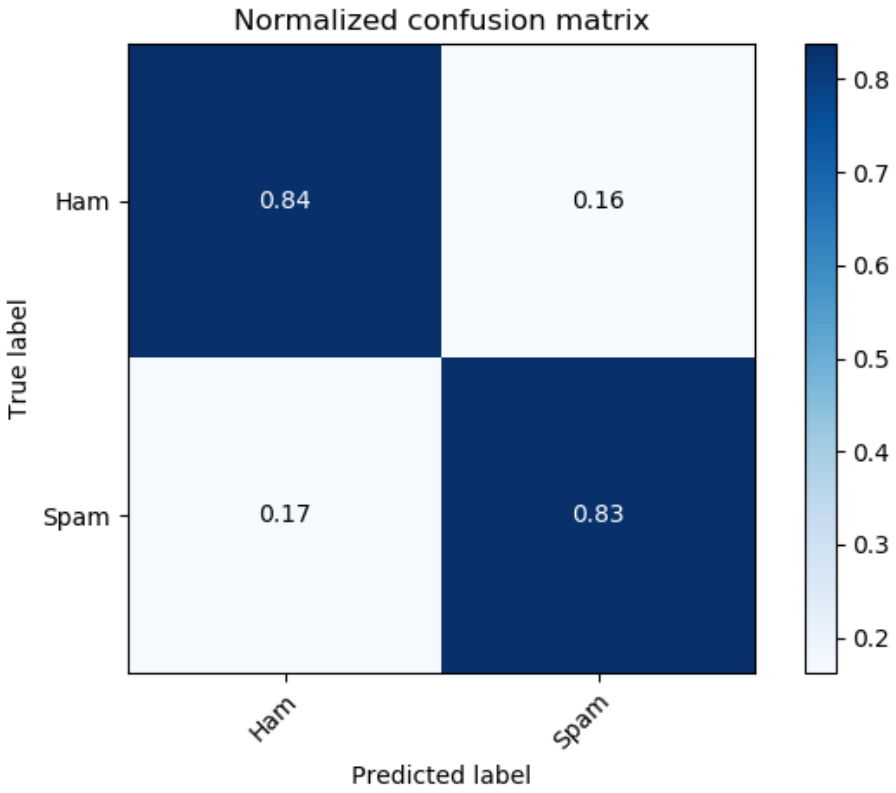
2.3 文本标题+时间（假设连续）

预处理：TF-IDF 处理文本标题，假设日期是周期性连续特征值，例如相邻的两天是连续的 24 小时周期，那么可以计算出该周期值的正弦和余弦值

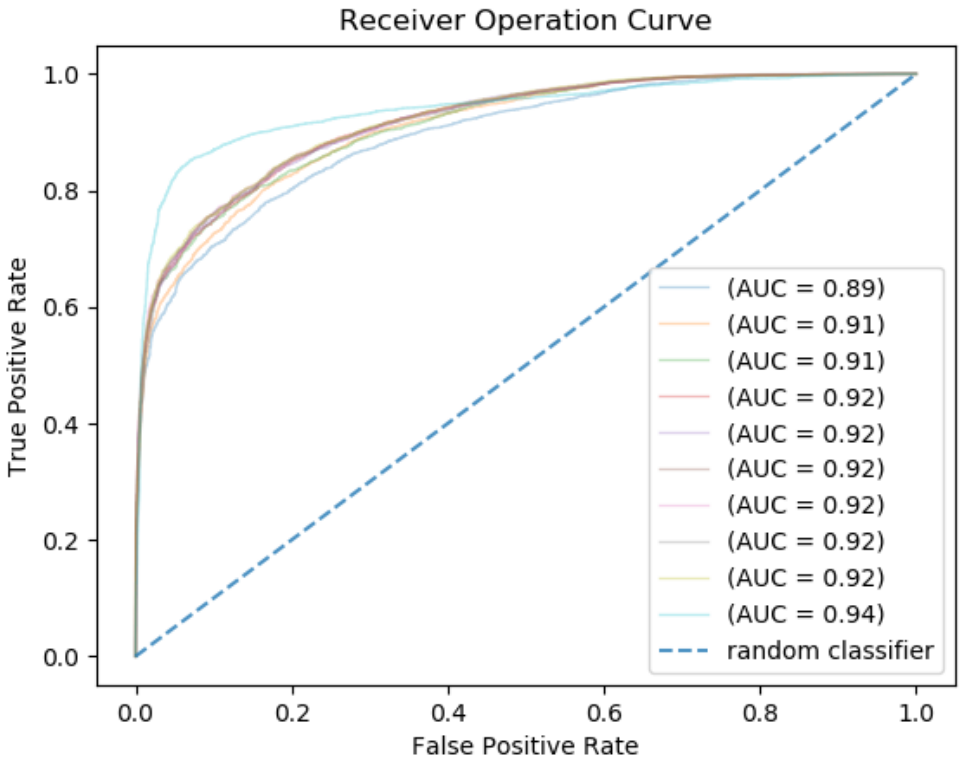
机学模型：Multinomial Naïve Bayes


Confusion matrix, without normalization
Ham
Spam
True label
7840
1522
1519
7554
Ham
Spam
Predicted label
7000
6000
5000
4000
3000
2000

图表 2.3.1 混淆矩阵



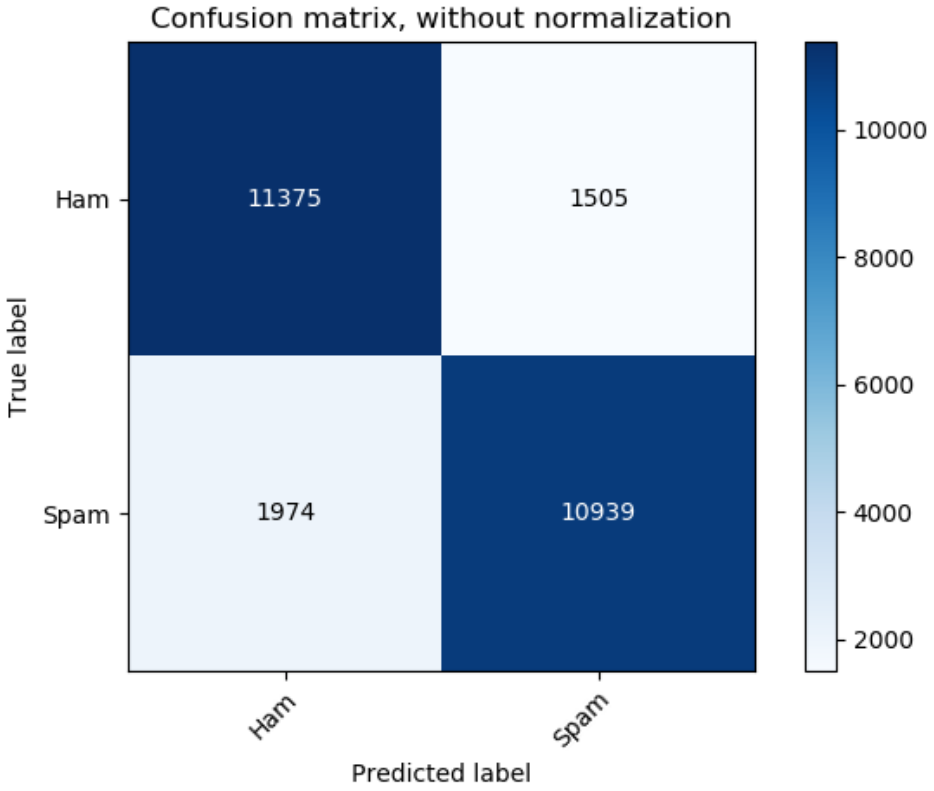
图表 2.3.2 标准化混淆矩阵

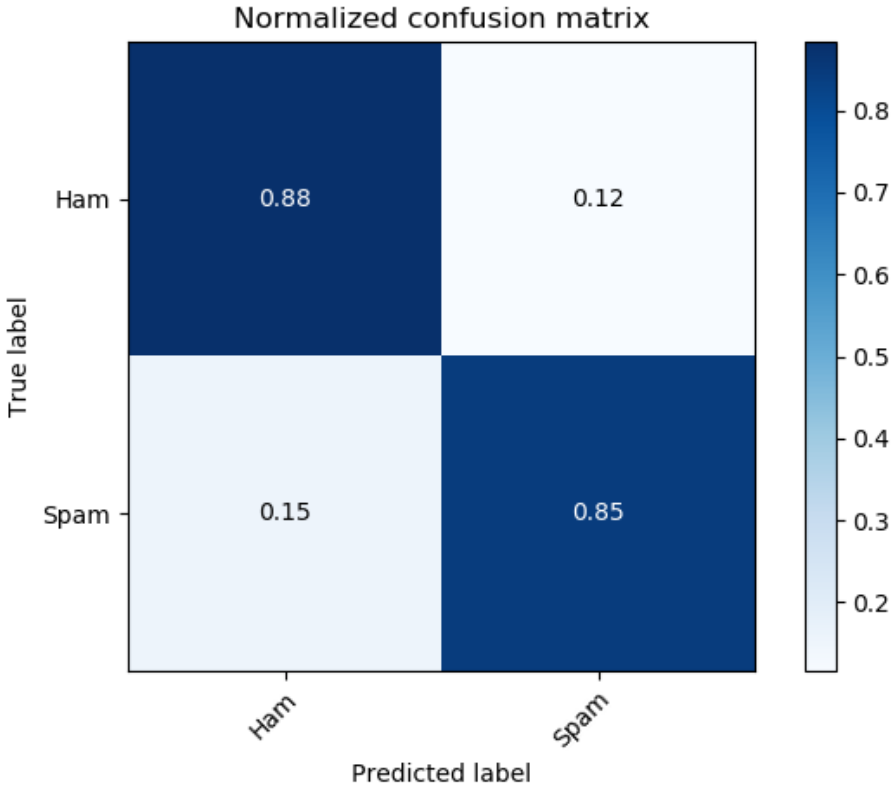


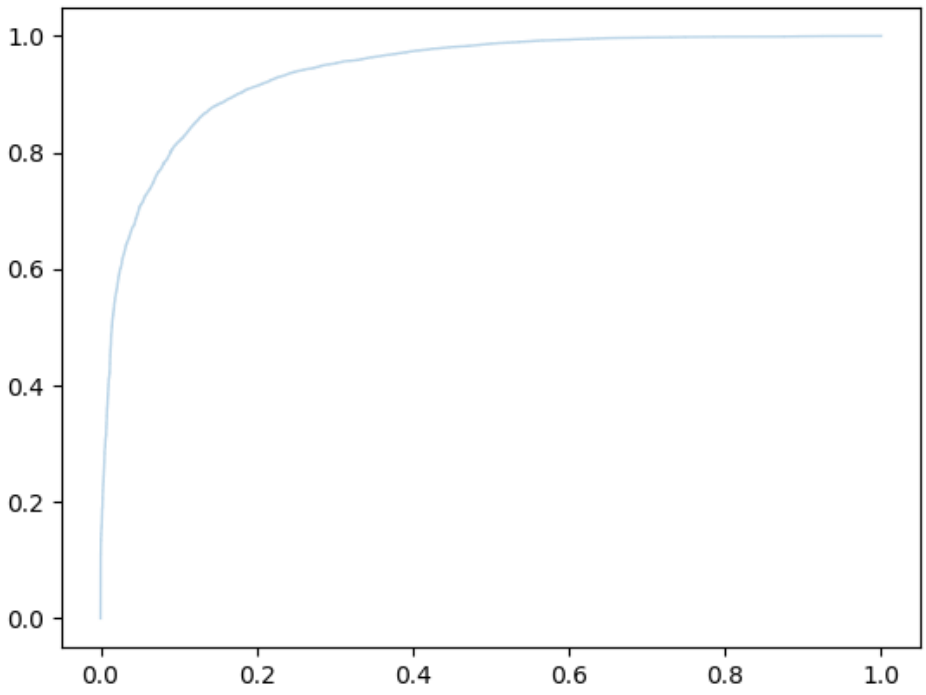
图表 2.3.3 10 折交叉验证 ROC 曲线

2.4 文本标题+日期（离散）+IP

预处理：词袋模型处理文本标题，离散处理日期数据，将 IPV4 地址转化为 32 位 2 进制字节码。另外，在该项训练中，我们保留了所有的缺失数据，并以缺失独热编码，例如以某个特定字段值 0 表示对于 IP 缺失的数据，1 表示存在，对缺失值采取补零操作，这样所有的数据都能用于模型的训练与测试。
机学模型：Bernoulli Naïve Bayes
评估图表：

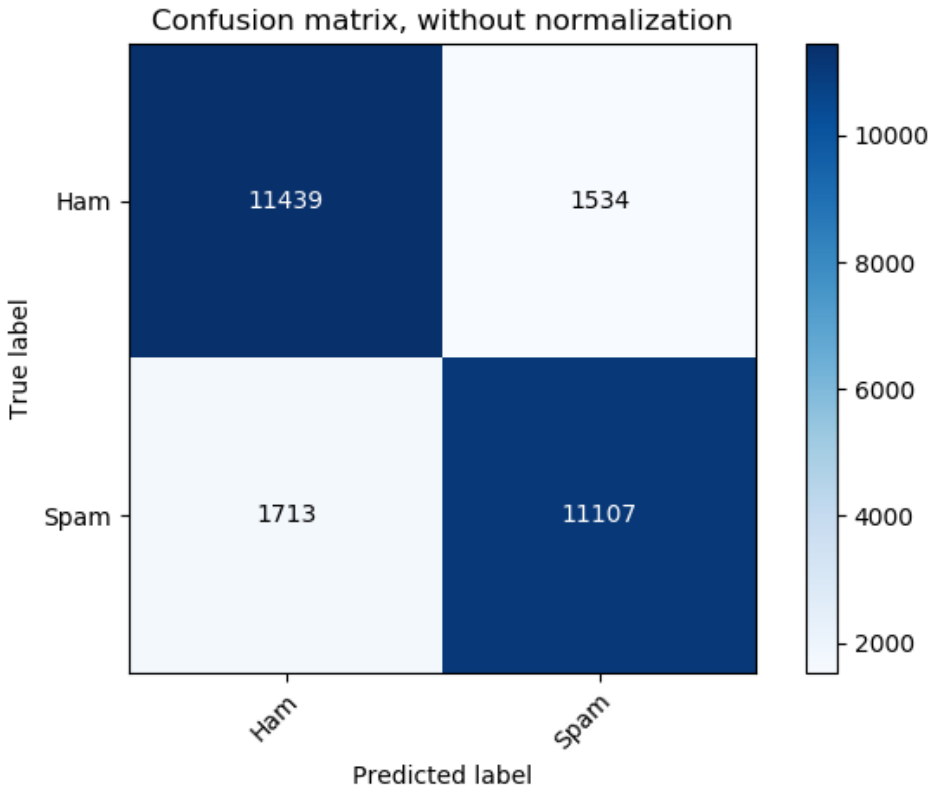


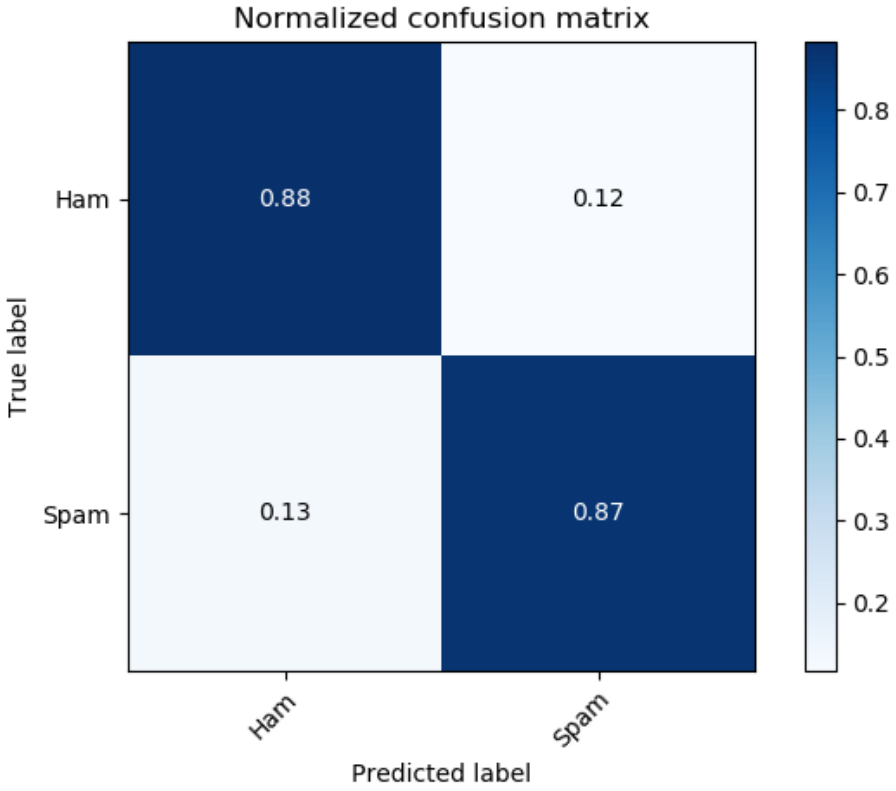

Normalized confusion matrix
Ham
Spam
True label
0.88
0.12
0.15
0.85
Ham
Spam
Predicted label
0.8
0.7
0.6
0.5
0.4
0.3
0.2

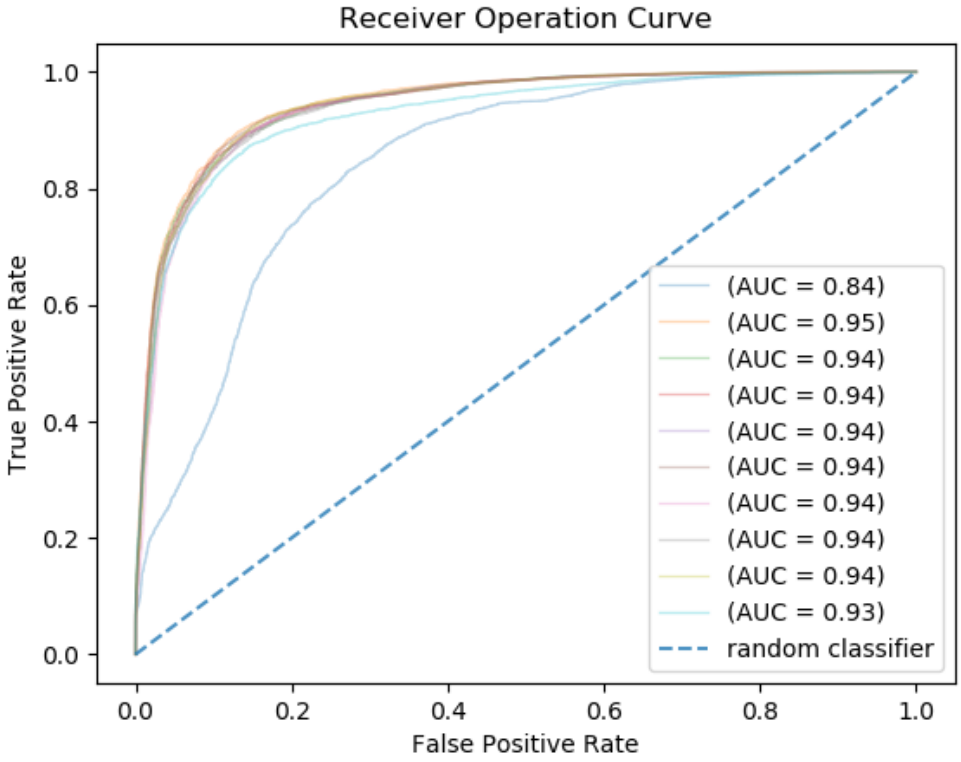

1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

2.5 文本标题+日期（离散）+ IP + 发件人邮箱 + 收件人 + 抄送人

预处理：词袋模型处理文本标题，离散处理日期数据，将 IPV4 地址转化为 32 位 2 进制字节码，发件人邮箱前后缀分别编码，前缀即用户名，从中提取大写字母、小写字母、特殊字符、数字各自占比情况，后缀即域名，提取常见的域名作为域名库，以独热编码，其它后缀以未知域名对待。统计收件人和抄送人的数量进行编码。
机学模型：Bernoulli Naïve Bayes
指标图表：




Normalized confusion matrix
Ham
Spam
True label
0.88
0.12
0.13
0.87
Ham
Spam
Predicted label
0.8
0.7
0.6
0.5
0.4
0.3
0.2


Receiver Operation Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
(AUC = 0.84)
(AUC = 0.95)
(AUC = 0.94)
(AUC = 0.94)
(AUC = 0.94)
(AUC = 0.94)
(AUC = 0.94)
(AUC = 0.94)
(AUC = 0.94)
(AUC = 0.93)
random classifier

三、总结

K 折交叉验证平均评估指标：

评估指标 / 特征	测试准确率	精准率	召回率	F1	AUC
文本标题	0.8080+/-0.0035	0.8053+/-0.0060	0.8181 +/-0.0045	0.8116 +/-0.0043	0.9016+/-0.0028
文本标题+日期(离散)	0.8556 +/-0.0130	0.8783+/-0.0059	0.8232+/-0.0353	0.8494+/-0.0177	0.9380+/-0.0097
文本标题+日期(连续)	0.8285+/-0.0167	0.8312+/-0.0092	0.8207+/-0.0294	0.8257+/-0.0188	0.9157+/-0.0105
文本标题+日期(离散)+IP	0.8567+/-0.0232	0.8635+/-0.0334	0.8492+/-0.0124	0.8560+/-0.0201	0.9303+/-0.0263
文本标题+日期(离散)+IP + 发件人邮箱地址+收件人+抄送人	0.8639+/-0.0296	0.8650+/-0.0416	0.8655 +/-0.0104	0.8648+/-0.0249	0.9315+/-0.0296

机器学习模型的测试与传统程序的测试不一样，传统程序，给定输入与输出，检测是否匹配。机器学习模型受训练数据的影响，训练数据量越大，越全面，模型学习到的知识越多和广泛，表现越优秀，输入与输出的匹配率越高，反之若一条知识并未出现在训练集，却出现在测试集，测试必然失败。我们采用 K 折交叉验证来测试机学黑盒子在不同训练数据下的稳定性。

K 折交叉验证：将所有数据平均分成 K 份，其中 K-1 份数据用来训练，1 份用来测试，并循环 K 次，这样保证了所有的数据都参与到了模型的训练与测试。上述该表的所有数据均设置 k=10，取 10 次训练下的平均指标与标准差。

将日期做连续特征值提取并没有帮助到模型，但是，日期离散特征提取对模型的各项指标均有一定的提升，测试准确度提升了 5%，精准度提升了 7%，但是召回率并没有大的提升。也就是说引入日期特征值，实际提升的是特异度（真阴性率）能让更多的正常邮件被成功预测，从而给垃圾邮件的预测度带来一定的提升，依然缺乏对垃圾邮件分类模型提升的关键数

据。后续,我们采取缺失值补位的方法，加大了数据量，并且尝试从邮件数据包括 IP、发件人、收件人中挖掘一些可能对分类有用的特征，对模型表现有一定的提升，在 1-3%左右，均衡了正负类的预测，最终各项指标稳定在 86.5%左右。