第七周学习汇报

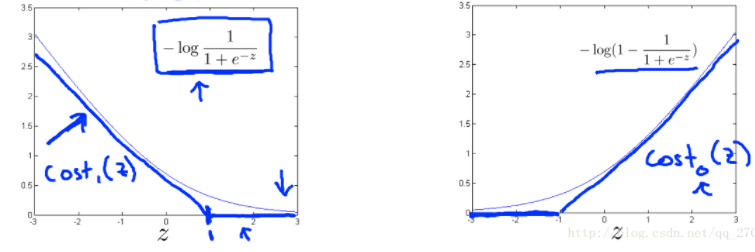
1. SVM代价函数

由逻辑回归代价函数：



进行改变得：

将逻辑回归代价函数的改变为蓝色折线部分。左图为y=1时的情况，右图为y=0时的情况。



所以有：y=1时，要求z大于等于1；y=0时，要求小于等于-1。

SVM的代价函数改变为以下形式：（乘以m，c相当于lamda，不影响最优值）



当，SVM代价函数最优化简化为：



即要求的最小值。



要想最小，则要求P最大。P为x在的投影，为的长度。

，从直观上感受SVM作为大间距分类器的效果。SVM也称为大间距分类器。

1. 核函数

核函数的预测函数：（构造新的特征向量）



是通过相似度函数得出：



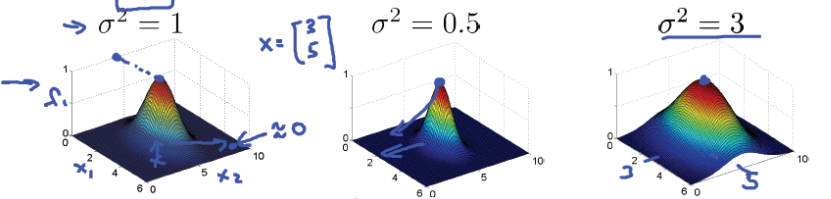
如果有，则；如果有使得很大，则。

当x离l很近的时候，相似度近似为1。

当x离l很远的时候，相似度近似为0。

指标记点，标记点的选择一般选择x的数据集；

的作用如下：



步骤为：

1. 选择标记点
2. 通过核函数计算相似度
3. 代入预测函数进行预测

对于m个训练数据，每个训练数据对应一个特征向量f中，总有一维向量的值为1。与之前的代价函数的区别在于用f代替了x。

对核函数进行归一化是必要的。

分类：

1. 线性核函数：
2. 高斯核函数：，通过把低维映射到高维

使用哪种算法进行预测？

1. 当n>m时，没有足够的样本进行预测，使用逻辑回归或者线性核函数的SVM。
2. 当m>n时，m属于中等大小时，使用高斯核函数的SVM。
3. 当m>n时，m属于特别大时，先增添特征向量，使用逻辑回归或者线性核函数的SVM。此时使用高斯核函数时，会显得特别慢。
4. 神经网络可以解决任何问题，但解决效率低。

编程作业：

1. 实现高斯核函数

利用高斯核函数公式计算相似度：

sim=exp(-(sum((x1-x2).^2))/(2\*sigma^2));

1. 训练得到C和sigma，在预先的集合中训练

cc = [0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30];

ss = cc;

maxx = 0;

for i=1:length(cc)

for j=1:length(cc)

model=svmTrain(X, y, cc(i), @(x1, x2) gaussianKernel(x1, x2, ss(j)));%提供的model训练函数

predictions=svmPredict(model, Xval);%提供的预测函数

error=mean(double(predictions == yval));%计算预测误差

if error<maxx %如果误差变小，则进行更新

maxx=error;

C=cc(i);

sigma=ss(j);

end

end

end

1. 对邮件字符串进行筛选

for i=1:length(vocabList)

if strcmp(vocabList{i}, str)%字符串比较

word\_indices = [word\_indices; i];%添加

break;

end

end

学习心得：

本周学习了SVM支持向量机算法，主要是通过把低维向高维转化解决非线性问题。SVM通过核函数（相似度函数）设置新的特征向量f更好的解决非线性问题。SVM算法主要是要选择C和sigma两个参数，通过训练得到合适的参数。通过编程作业实现高斯核函数，训练得到C和sigma参数，并实践垃圾邮件分类。