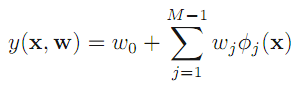
# 3 Linear Models for Regression—PRML

## 3.1 Linear Basis Function Models



There are D features.

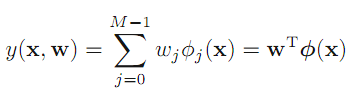


Where  are knows as basis function.

例如： 

可以通过，

转化为求解线性回归问题：



, 

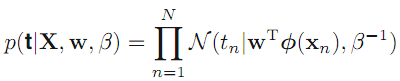
Maximum likelihood and least squares

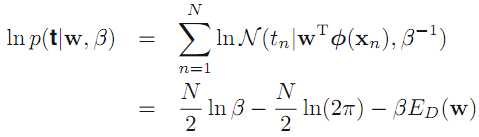
Assume  where is a 0 mean Gaussian random variable with variance .

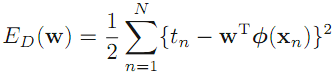
For a sample (x,t)



For training dataset (n samples) :





Minmize 

### Batch learning

矩阵求导的基本知识：







### Sequential learning

采用随机梯度法避免一次载入所有学习样本，，即：

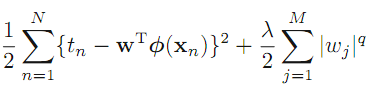
，（矩阵表示），

（普通表示）。

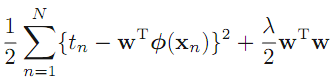
（先设w=[0,0,0,…]T,随机选一个样本计算，并更新w，直到前后两次w的差小于某个阈值）

### Regularized least squares

防止过拟合，采用正则化，保证选择出一个既与学习数据相近的，而且最简单（参数最少）的模型。



当q=1时：lasso。

当q=2时，Ridge regression。，









因此堆学习的解为：



而序列学习的解为：



### Multiple outputs

如果输出为3元组，对于N个学习样本，记，所有公式用T替代t即可。相当于对输出三元组的每一维进行回归，计算对应的w，则应当有3组w。

## 3.2 The Bias-Variance Decomposition

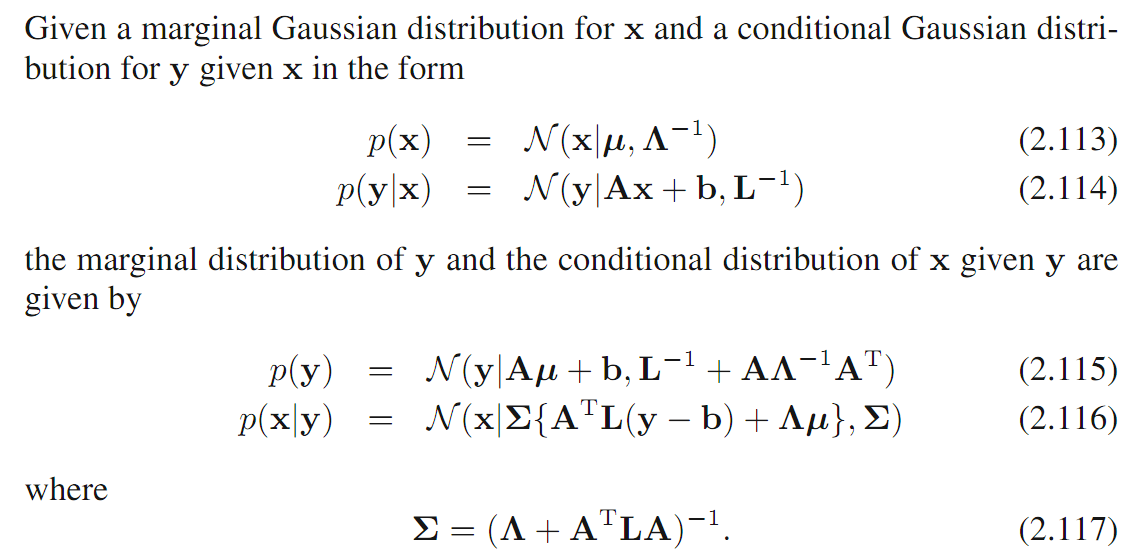
D为训练样本集合，当D不同时，学习得到的h(\*)不同，假设考虑预测某一个样本x对应的y，则根据不同学习样本D得到的预测值h(x)的期望误差为

解释：当估计的参数越多时，Bias越小（参考泰勒展开），但是当学习样本集发生变化时，参数多的h变化也大，因此var(h)也会增大。

## 3.3. Bayesian Linear Regression

设：，

且关于高斯分布参数的共轭分布仍为高斯分布，即：



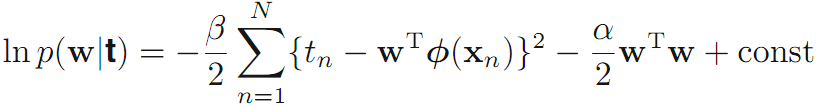
则：



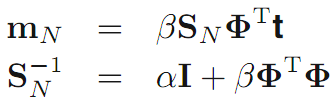
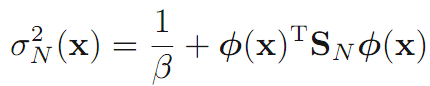
两种解法：

1. 选择最大的W，并计算近似。
2. 计算精确解。

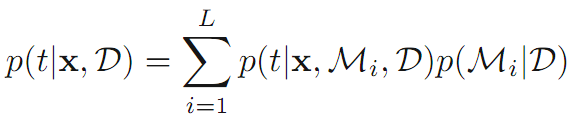
第一种：得到与ridge regression一致的结果。

即求解w，使得最大。

第二种：根据共轭发布的性质，可得到结果为：，

where，

对于模型选择来说，贝叶斯方法可以理解为将所有可能的模型预测结果进行加权，因此完全贝叶斯方法可以避免过拟合，难点在于常常难以计算出后验概率和模型平均。



## Local Linear Regression

根据n个样本(xn，tn)，估计x对应的t

，



求解，即对w求导：





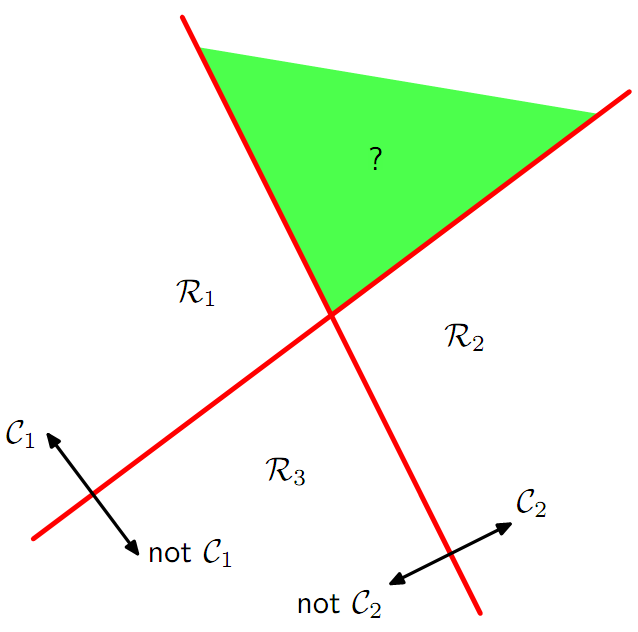
根据预测的点不同，每次计算得到的w都不同，即非参数估计。

# 4 Linear Models for Classification

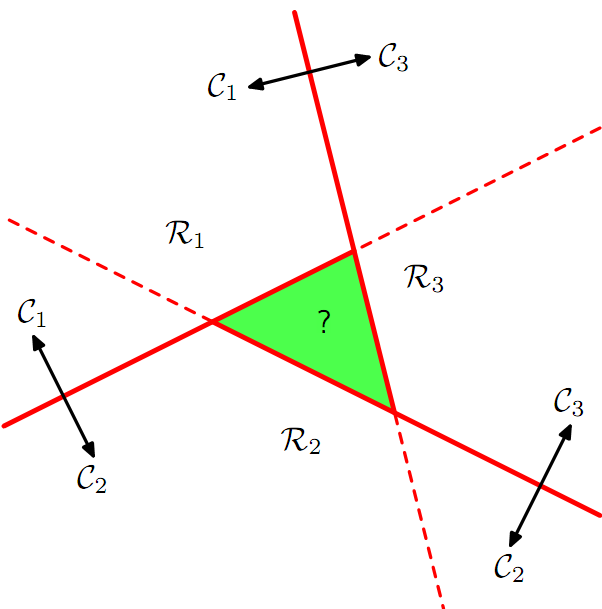
## 4.1Discriminant Functions



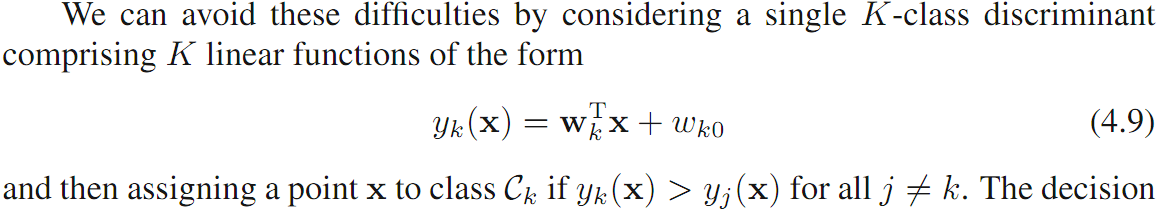
4.1.1 Two classes



4.1.2 Multiple classes



通过为每个类建立判别函数实现分类，可以避免以上问题：



### 4.1.3 Least squares for classiﬁcation

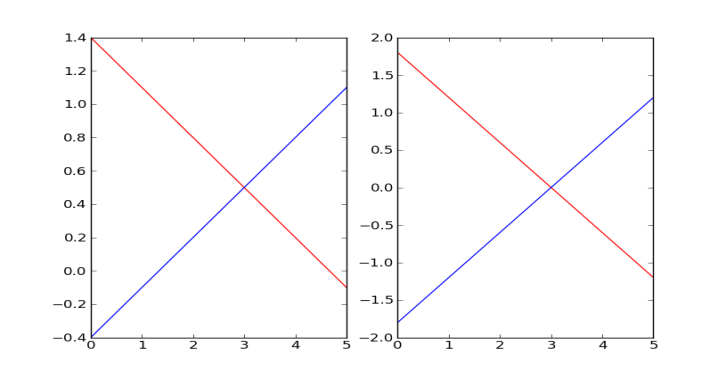
例如样本点为(1,y1),(2,y1),(4,y2),(5,y2)，将类别标签写为y=（1,0）if y=y1 otherwise (0,1)

进行最小二乘法计算Y=WX中的W

根据最小二乘法公式： W=(XXT)-1XY，，，

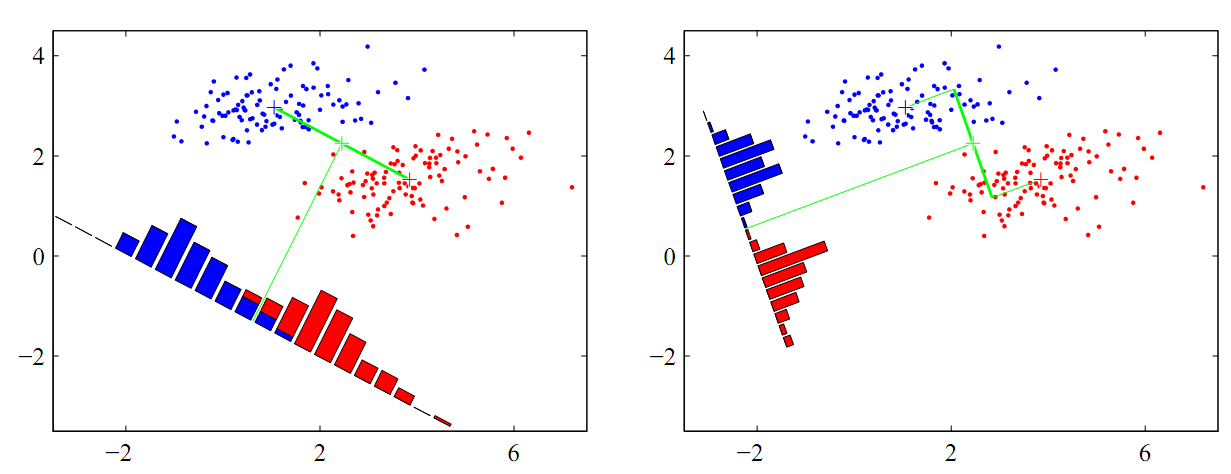
计算y1的判别函数：W=(XXT)-1XY1=[1.4,-0.3]: y = 1.4 – 0.3x 如果用1和-1表示类别，(1.8,-0.6)

计算y2的判别函数：W=(XXT)-1XY2=[-0.4,0.3]: y = -0.4 + 0.3x 如果用1和-1表示类别，(-1.8,0.6)



### 4.1.4 Fisher’s linear discriminant

选择一个方向投影，使得两个类的投影距离最大，且投影后每个类内部的方差最小。（因为类内部方差越大，两个类投影后之间就越可能有重叠，如下图）



求投影方向，使得两个类的平均点投影距离最大，且投影后每个类内部的方差最小。

设为类别K的所有样本的均值，

两个类的平均点投影距离：



第一个类投影后的内部方案为：



Fisher判别的目标函数为：

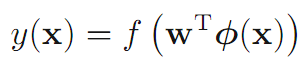
Max，且

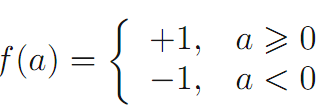
对W求导，得到解为：



根据样本选择阈值y0，对于未知样本x，wx与y0比较后实现分类。

### 4.1.5 The perceptron algorithm

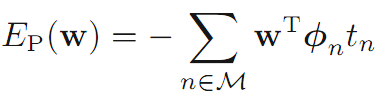
二类分类问题：判别函数，其中，

，例如：

设类别t取值为-1，1.

则对于分类正确的标准可转换为：连续函数（大多数分类算法的trick）

则目标函数为

最小化，M表示错误分类的样本。

使用随机梯度下降法SGD：



**即对每个样本x,采用当前w计算，预测为-1类或1类（），如果分类错误，则更新w得到w’，重复直到收敛。**

## 4.2 Probabilistic Generative Models

设：x的分布函数为多维高斯分布，D为维度。各类的协方差相等都为



考虑二类分类问题：



即样本计算出的似然大于0时为第一类，小于0为第二类等于0不可分。

上式为x的线性函数，因此叫做线性判定准则（Linear Discriminant Analysis）

如果各类的协方差不相等，为，



上式为x的二次函数，因此叫做二次判定准则（Quadratic Discriminant Analysis）

**如果为指数族分布函数**

**（例如多项式分布，高斯分布，指数分布等），那么计算出似然函数L(X)可以发现，他们均为线性函数WX+wo的形式**

## 4.3 Probabilistic Discriminative Models

### Logistic Regression





求解可以通过最大似然估计，即对于样本xi，标签ti为0，1.



即求解W使得

采用牛顿法迭代求解：设

，



不断迭代求解



### Softmax regression

Givenatestinput*x*,wewantourhypothesistoestimatetheprobabilitythat*p*(*y*=*j*|*x*)foreachvalueofj = 1, \ldots, k.I.e.,wewanttoestimatetheprobabilityoftheclasslabeltakingoneachofthe*k*differentpossiblevalues.Thus,ourhypothesiswilloutputa*k*dimensionalvector(whoseelementssumto1)givingusour*k*estimatedprobabilities.Concretely,ourhypothesis*h*θ(*x*)takestheform:


\begin{align}
h_\theta(x^{(i)}) =
\begin{bmatrix}
p(y^{(i)} = 1 | x^{(i)}; \theta) \\
p(y^{(i)} = 2 | x^{(i)}; \theta) \\
\vdots \\
p(y^{(i)} = k | x^{(i)}; \theta)
\end{bmatrix}
=
\frac{1}{ \sum_{j=1}^{k}{e^{ \theta_j^T x^{(i)} }} }
\begin{bmatrix}
e^{ \theta_1^T x^{(i)} } \\
e^{ \theta_2^T x^{(i)} } \\
\vdots \\
e^{ \theta_k^T x^{(i)} } \\
\end{bmatrix}
\end{align}


For the softmax regression, If we subtract a fix vector ψfromeveryθ*j*, it doesnotaffectourhypothesis'predictionsatall!See the explaining as following:


\begin{align}
p(y^{(i)} = j | x^{(i)} ; \theta)
&= \frac{e^{(\theta_j-\psi)^T x^{(i)}}}{\sum_{l=1}^k e^{ (\theta_l-\psi)^T x^{(i)}}}  \\
&= \frac{e^{\theta_j^T x^{(i)}} e^{-\psi^Tx^{(i)}}}{\sum_{l=1}^k e^{\theta_l^T x^{(i)}} e^{-\psi^Tx^{(i)}}} \\
&= \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^k e^{ \theta_l^T x^{(i)}}}.
\end{align}


Inpractice,however,itisoftencleanerandsimplertoimplementtheversionwhichkeepsalltheparameters(\theta_1, \theta_2,\ldots, \theta_n),withoutarbitrarilysettingoneofthemtozero.Butwewillmakeonechangetothecostfunction:Addingweightdecay.Thiswilltakecareofthenumericalproblemsassociatedwithsoftmaxregression'soverparameterizedrepresentation.

Wewillmodifythecostfunctionbyaddingaweightdecayterm\textstyle \frac{\lambda}{2} \sum_{i=1}^k \sum_{j=0}^{n} \theta_{ij}^2whichpenalizeslargevaluesoftheparameters.Ourcostfunctionisnow


\begin{align}
J(\theta) = - \frac{1}{m} \left[ \sum_{i=1}^{m} \sum_{j=1}^{k} 1\left\{y^{(i)} = j\right\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^k e^{ \theta_l^T x^{(i)} }}  \right]
              + \frac{\lambda}{2} \sum_{i=1}^k \sum_{j=0}^n \theta_{ij}^2
\end{align}


Withthisweightdecayterm(foranyλ>0),thecostfunction*J*(θ)isnowstrictlyconvex,andisguaranteedtohaveauniquesolution.TheHessianisnowinvertible,andbecause*J*(θ)isconvex,algorithmssuchasgradientdescent,L-BFGS,etc.areguaranteedtoconvergetotheglobalminimum.

Toapplyanoptimizationalgorithm,wealsoneedthederivativeofthisnewdefinitionof*J*(θ).Onecanshowthatthederivativeis:


\begin{align}
\nabla_{\theta_j} J(\theta) = - \frac{1}{m} \sum_{i=1}^{m}{ \left[ x^{(i)} ( 1\{ y^{(i)} = j\}  - p(y^{(i)} = j | x^{(i)}; \theta) ) \right]  } + \lambda \theta_j
\end{align}


Inthespecialcasewhere*k*=2,onecanshowthatsoftmaxregressionreducestologisticregression.Thisshowsthatsoftmaxregressionisageneralizationoflogisticregression.Concretely,when*k*=2,thesoftmaxregressionhypothesisoutputs


\begin{align}
h_\theta(x) &=

\frac{1}{ e^{\theta_1^Tx}  + e^{ \theta_2^T x^{(i)} } }
\begin{bmatrix}
e^{ \theta_1^T x } \\
e^{ \theta_2^T x }
\end{bmatrix}
\end{align}


Takingadvantageofthefactthatthishypothesisisoverparameterizedandsettingψ=θ1,wecansubtractθ1fromeachofthetwoparameters,givingus


\begin{align}
h(x) &=

\frac{1}{ e^{\vec{0}^Tx}  + e^{ (\theta_2-\theta_1)^T x^{(i)} } }
\begin{bmatrix}
e^{ \vec{0}^T x } \\
e^{ (\theta_2-\theta_1)^T x }
\end{bmatrix} \\

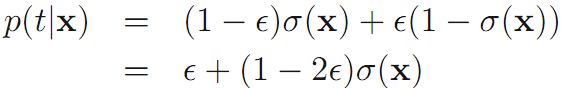

&=
\begin{bmatrix}
\frac{1}{ 1 + e^{ (\theta_2-\theta_1)^T x^{(i)} } } \\
\frac{e^{ (\theta_2-\theta_1)^T x }}{ 1 + e^{ (\theta_2-\theta_1)^T x^{(i)} } }
\end{bmatrix} \\

&=
\begin{bmatrix}
\frac{1}{ 1  + e^{ (\theta_2-\theta_1)^T x^{(i)} } } \\
1 - \frac{1}{ 1  + e^{ (\theta_2-\theta_1)^T x^{(i)} } } \\
\end{bmatrix}
\end{align}


Thus,replacingθ2−θ1withasingleparametervectorθ',wefindthatsoftmaxregressionpredictstheprobabilityofoneoftheclassesas\frac{1}{ 1  + e^{ (\theta')^T x^{(i)} } },andthatoftheotherclassas1 - \frac{1}{ 1 + e^{ (\theta')^T x^{(i)} } },sameaslogisticregression.

LR回归会因为离群点或错样本受到影响，如果存在大量错误样本，可以通过以下办法处理：

设置超参数，表示该样本是错误样本的概率，则：



## 4.4 Laplace Approximation

对于随机变量x，概率密度函数f(x)的归一化形式为，如果采用高斯分布逼近该函数，则称为Laplace Approximation。即：



其中，A是ln f(x)的Hessian矩阵在极值点z上的取值，即  


### BIC、AIC、HQC

模型选择标准之一： \mathrm{BIC} = {-2 \cdot \ln{\hat L} + k \cdot \ln(n)}. \ ，BIC越小越好。

缺点：

1. n必须远远大于m；
2. 对高维问题中的特征选择模型效果不好

**类似的有\mathrm{AIC} = 2k - 2\ln(L)、\mathrm{AICc} = \mathrm{AIC} + \frac{2k(k + 1)}{n - k - 1}、 \mathrm{HQC} = -2 L_{max} + 2 k \log \log n, \ **

## 4.5. Bayesian Logistic Regression

设，其中W的先验概率服从多维高斯分布（其中为超参数），对于学习样本集合T，求新样本x, t’的分类结果。



其中，，, **,**为后验概率的w最大似然估计。





采用Laplace估计可以得到的高斯估计：

其中z0为w的最大似然估计，

****

其中，，