1. 要证明n个训练数据点 $z_1 \cdots z_n$ 分空间的 Voronoi 网格为凸，只需证明 $\forall z_i, z_j \in \{z_1 \cdots z_n\}$

离 z_i 比 z_j 更近的点构成一个凸集 $S_{i>j}$ 即可.

因为 S_i 为最近的 Voronoi 网格单满足 $S_i = \bigcap_{j=1, j\neq i}^{j=n} S_{i>j}$ 若 $\forall S_{i>j}$ 为凸 那么 S_i 由集合交运算保凸性知为凸。

对于两个 z_i, z_j，$S_{i>j}$ 可写为 $\{x \mid \|x - z_i\| \leq \|x - z_j\|\}$

当 $\|\cdot\|$ 为欧式距离时 有 $x^T x - z_i^T x - x^T z_i + z_i^T z_i \leq x^T x - z_j^T x - x^T z_j + z_j^T z_j$

$$\Rightarrow 2(z_j - z_i)^T x - (z_j^T z_j - z_i^T z_i) \leq 0$$

$S_{i>j}$ 为一个 hyperspace 为凸！得证.

2. 1) 贝叶斯误差率 $e = \sum_i \int_{x \in w_i 判决区} p(w_i) p(x|w_i)$

由于在 $x \in [0, \frac{cr}{c-1}]$ 区域所有的似然一样，先验一样，判决都可行

不妨设为 w_j 判决域，在 $x \in [i, i+1-\frac{cr}{c-1}]$ 时不会出现判决失误 $\forall i \in \{1, \cdots\}$

有 $e = \sum_{i=1}^{c} \frac{r}{c} \frac{cr}{c-1}$

$= r$

2) 最近邻规则相当于若 $x \in [0, \frac{cr}{c-1}]$ 则判决为 w_1，其它情况与贝叶斯判决一样

据1)中推导显然 错误率为贝叶斯误差率.

3. 距离度量需满足 1. 非负性： $D_m(x,y) = [\sum_{i=1}^{d} |x_i - y_i|^s]^{\frac{1}{s}} > 0$ 明显满足

2. 对称性 $D_m(x,y) = D_m(y,x)$ 显然

3. 自己与自己距离为0 $D_m(x,x) = 0$ 显然

4. 三角不等式： $D_m(x,y) + D_m(y,z) \geq D_m(x,z)$

~~运用：(等式放大) 见下页~~

~~($D_m(x,y) + D_m(y,z)$) ...~~

$$D_m(x,z)^s = \sum_{i=1}^{d} (x_i - z_i)^s$$

$$= \sum_{i=1}^{d} |x_i - y_i + y_i - z_i| |x_i - z_i|^{s-1}$$

由于 $|x_i - z_i| = |x_i - y_i + y_i - z_i| \le |x_i - y_i| + |z_i - y_i|$

$$D_m(x,z)^s \le \sum_{i=1}^{d} |x_i - y_i| |x_i - z_i|^{s-1} + \sum_{i=1}^{d} |z_i - y_i| |x_i - z_i|^{s-1}$$

用赫尔德不等式 $\frac{1}{t} + \frac{1}{s} = 1$

$$\le \left(\sum_{i=1}^{d} |x_i - y_i|^s\right)^{\frac{1}{s}} \left[\sum_{i=1}^{d} (x_i - z_i)^{t(s-1)}\right]^{\frac{1}{t}}$$

$$+ \left(\sum_{i=1}^{d} |z_i - y_i|^s\right)^{\frac{1}{s}} \left[\sum_{i=1}^{d} (x_i - z_i)^{t(s-1)}\right]^{\frac{1}{t}}$$

$$= \left[\left(\sum_{i=1}^{d} |x_i - y_i|^s\right)^{\frac{1}{s}} + \left(\sum_{i=1}^{d} |z_i - y_i|^s\right)^{\frac{1}{s}}\right] \left[\sum_{i=1}^{d} |x_i - z_i|^{t(s-1)}\right]^{\frac{1}{st}}$$

$$= \left[D_m(x,y) + D_m(z,y)\right] \left[D_m(x,z)\right]^{\frac{s}{t}}$$

$$\Rightarrow D_m(x,z)^{s-\frac{s}{t}} \le D_m(x,y) + D_m(z,y)$$

$$s - \frac{s}{t} = 1$$

$$\Rightarrow D_m(x,z) \le D_m(x,y) + D_m(z,y)$$

三角不等式

4. 1) 首先 $p(x|\beta^{-1},q) > 0$ 显然

归一性：$\int_{-\infty}^{+\infty} p(x|\beta^{-1},q) = 2\int_0^{+\infty} p(x|\beta^{-1},q)dx = q(\frac{\beta}{2})^{\frac{1}{q}} \frac{1}{\Gamma(\frac{1}{q})} \int_0^{+\infty} \exp(-\frac{\beta}{2}x^q)dx$

分部积分

$$\stackrel{\text{分部积分}}{=} q(\frac{\beta}{2})^{\frac{1}{q}} \frac{1}{\Gamma(\frac{1}{q})} \left[x\exp(-\frac{\beta}{2}x^q)\Big|_0^{\infty} + \frac{\beta}{2}\int_0^{+\infty} x\exp(-\frac{\beta}{2}x^q)dx^q \right]$$

换元 $z = x^q$

$$= q(\frac{\beta}{2})^{\frac{1}{q}} \frac{1}{\Gamma(\frac{1}{q})} \cdot \frac{\beta}{2}\int_0^{+\infty} z^{\frac{1}{q}} \exp(-\frac{\beta}{2}z)dz$$

$$\stackrel{t=\frac{\beta}{2}z}{=} q(\frac{\beta}{2})^{\frac{1}{q}} \frac{1}{\Gamma(\frac{1}{q})} \frac{\beta}{2}\int_0^{+\infty} (\frac{2}{\beta})^{\frac{1}{q}} t^{\frac{1}{q}} \exp(-t)\frac{2}{\beta}dt$$

$$= \frac{q}{\Gamma(\frac{1}{q})}\int_0^{+\infty} t^{\frac{1}{q}}\exp(-t)dt = \frac{q}{\Gamma(\frac{1}{q})}\Gamma(\frac{1}{q}+1) = q\cdot\frac{1}{q} = 1$$

得证

2) 似然函数为

$$L = \prod_{i=1}^{N} \frac{q}{2}(\frac{\beta}{2})^{\frac{1}{q}} \frac{1}{\Gamma(\frac{1}{q})} \exp(-\frac{\beta}{2}|y(x_i,w) - t_i|^q)$$

对数似然 $\log L = \sum_{i=1}^{N}(-\frac{\beta}{2}|y(x_i,w) - t_i|^q) + Const$

则 $\log L = -\frac{\beta}{2}E(w) + Const$

即最大化 $\log L$ 的 MLE 为最小化 $E(w)$ 的优化

3) MAP 估计 $p(w|\{t,y\};\alpha,\beta)$ 的分子部分为 $p(w|\alpha)\cdot p(\{t,y\}|w)$

$$\log p(w|\alpha)p(\{t,y\}|w) = \log L + \log p(w|\alpha)$$

$$= -\frac{\beta}{2}E(w) - \frac{\alpha}{2}\sum_{j=1}^{M}|w_j|^p + const$$

令 $\lambda = \frac{\alpha}{\beta}$ 则有极大除了 const 项以外

与 $E'(w) = \sum_{n=1}^{N}|y(x_n,w) - t_n|^q + \lambda\sum_{j=1}^{M}|w_j|^p$ 只差负常数倍

得证 □

1. 训练样本对结果的影响

固定 K=10，每个数字类别取 10, 100, 1000， 各类样本数量的最小值= 5421 个训练样本得到的结果为:

N	空间复杂度	运算时间(s)	Error rate
10	O(N)	2.733832	0.3917
100		20.664020	0.1496
1000		134.686047	0.0577
5421		664.444613	0.0343

可以看到固定 K，随着每种数字类别的训练样本 N 的增加，测试分类错误率在减小，空间复杂度和时间复杂度基本上也成 O(N)增加。这符合我们的直观。

2. K 近邻的 K 对结果的影响

固定每类别取 1000 个，取 K 为 1, 2, 3, 4, 5, 10, 100, 1000 进行实验得到结果：

K	Error rate
1	0.0534
2	0.0653
3	0.0539
4	0.0564
5	0.0558
10	0.0577
100	0.1044
1000	0.2666

可以看到，在 N 只取 1000 的情况下，K=1 是 test 误差最小的。由于训练样本较少，K 越来越大时，error rate 反而越来越大，这应该是因为有不少的测试样本虽然很相近的点仍然为自己同类别的点，但是在取 K 较大时，反而会引入不少其他类别的点。极限情况：在 K=10000, N=1000 时已经相当于是瞎猜。

3. 距离量度对结果的影响

考虑 L3 距离，即 Minkowski 中 s=3 的情况。取 K=1,2,3,4 ，N=1000 得到结果如下：

K	Error rate
1	0.0487
2	0.0609
3	0.0495
4	0.0532

可以看到 L3 距离量度的准确率比 L2 距离量度的准确率要稍高一些。

4. Per-feature 加权对结果的影响

对每维 feature(即每个像素)赋予不同的权值的含义是，对于分类更重要的 feature 应该有更大一点的 weight。若从 feature 在类间、类内方差的角度来考虑 feature 对分类问题的重要性：对于第 k 维 feature，考虑两类 i, j 间 variance 和类内点的 variance 为:

$$S_{inter} = E\left[\left(x_{ik} - x_{jk}\right)^2\right] = E[x_{ik}^2] + E\left[x_{jk}\right]^2 - 2E[x_{ik}]E\left[x_{jk}\right]$$

$$= Var(x_{ik}) + Var\left(x_{jk}\right) + \left(m_{ik} - m_{jk}\right)^2$$

其中$m_{ik} = E[x_{ik}]$和$m_j = E[x_{jk}]$。假设(i,j)是从二分类 pair 中(一共$\binom{10}{2}$)中均匀概率选出。

$$E^{(k)}{}_{(i,j)\in PAIR}\left[S_{inter\ ij}\right] = \frac{2}{10-1}\sum_i Var(x_{ik}) + 2Var(m_k)$$

其中$2Var(m_k)$类间的 variance 期望为(设每类概率 1/10):

$$E^{(k)}[S_{intra}] = E^{(k)}_{i=0\ldots9}\left[E_{n,m\text{为第}i\text{类里两个点}}[x_{ikm} - x_{ikn}]\right] = \sum_i \frac{2}{10} Var(x_{ik})$$

十分直观的，如果第 k 维 feature 类间方差较大，类内方差较小，那么应该给较大的权重。这样的方式会使得对各类分类都不太重要的像素(比如最边缘的像素)上的噪声或扰动对分类结果影响更小。所以可以构造类似

$$w_k = \left(\frac{E^{(k)}{}_{(i,j)\in PAIR}\left[S_{inter\ ij}\right]}{E^{(k)}[S_{intra}] + \lambda}\right)^p$$

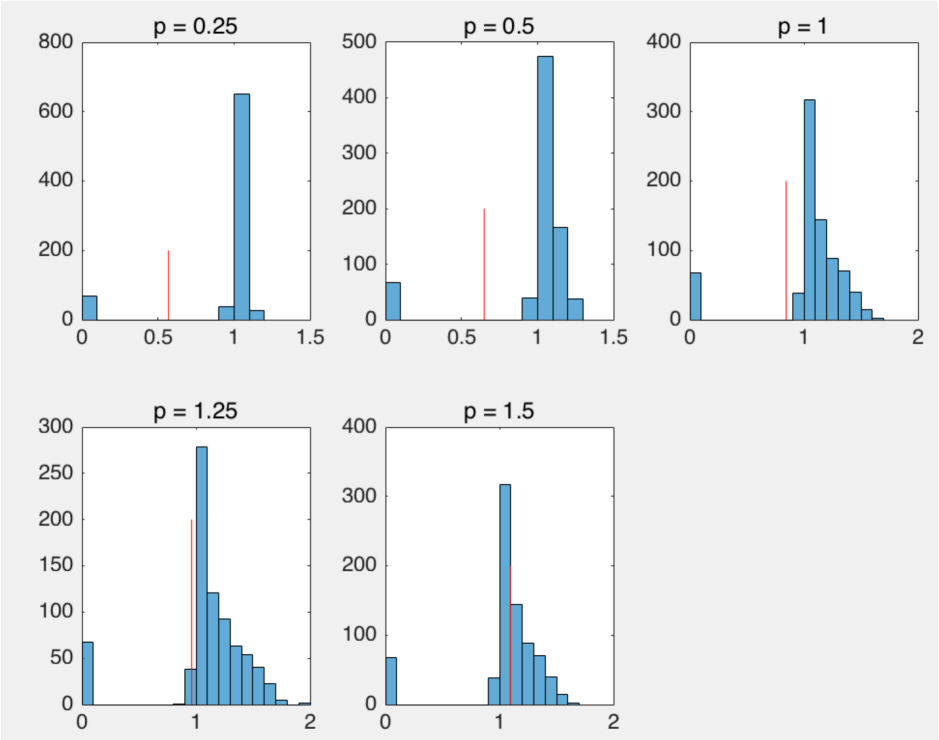
作为第 k 维 feature 的权重。其中λ为防止在 intra-class 方差期望太小时，而实际 inter-class 该 feature 也没有太大意义(inter-class 方差也很小)时， w_k极大甚至为 NAN 的情况，在我的实验中取$\lambda = 1e-3$。$p \geq 0$为调整这个 weight 影响的系数,当 p=0 时相当于没有调整原始 feature。

对 L3 距离, N=1000, K=1 这个我的电脑配置还能跑出来的数据里最好的结果上，加上以上预处理得到以下结果(其中 effective feature 在下面解释)：

p	Error rate	Effective feature(10)	Effective Feature(8)
0(原始 feature)	0.0487	784	784
0.25	0.0481	717	717
0.5	0.0472	717	717
1	0.0456	717	717
1.25	0.0468	717	715
1.5	0.0496	531	421

从上面几组实验看起来在$p \leq 1$时，的确越放大在类间类内差距比更大的 feature 的权重是有好处的。但是很明显这个 p 不能放到很大，因为 p 如果太大，有效的 feature 维数在降低，有很多虽然不那么重要但也能起作用的 feature 更多的被忽略了，所以 p 的调整存在一个重视 feature 之间的重要性和有效 feature 的 trade-off。为了衡量 effective feature 对 p 调整的指导作用，我设置了一个$th = \frac{1}{stage}^{\frac{1}{3}}$，代表在 L3 norm 时，如果$\frac{w_i}{w_j} < th$代表 feature i 差距是另一个 feature j 差距的 stage 倍时才能对结果起到一样的作用， 此时称 feature i 相对 feature j 无效。定义 feature 是 effective 的，只要 feature 相对于最优 feature 是 effective。

用这个方法来衡量不同 p 取值时有多少 effective feature。这里面 stage 是一个超参数，代表我认为对于手写数字识别，我不关心像素值有很小的波动，只关心 0~255 中的 stage 个像素等级。从手写字识别的这个问题中，作为人类，我们可以感觉出关心的灰度等级大概也就 8~10 个。下面画出不同的 p 取值 weight 的分布：其中红线为我们认为关心的灰度等级为 8 的时候，effective 和非 effective 的 weight 的分界值。从实际 p 取不同值的结果能看出的确在这个 task 中真正有效果的像素等级个数 stage 略小于 8(从第三张图可以看出，还需要把 threshold 往右移动一点)。



以下为这部分的实验截图：

```
>> main
power: 3; number of points per class: 1000; K: 1: error rate: 0.048700
>>
```

```
>> main
preprocess data: p: 0.250000
power: 3; number of points per class: 1000; K: 1: error rate: 0.048100
```

```
>> main
preprocess data: p: 0.500000
power: 3; number of points per class: 1000; K: 1: error rate: 0.047200
>> main
preprocess data: p: 1.000000
power: 3; number of points per class: 1000; K: 1: error rate: 0.045600
```

5. 切线距离实验

考虑平移操作(沿着图像水平/垂直方向平移)和旋转操作，这样的 transform 对应

的切线距离可以消除一些数据间微小的水平 / 垂直平移和微小旋转带来的影响。

垂直和水平平移变换的切向量 T 分别用$\begin{bmatrix}-1\\0\\1\end{bmatrix}$和[-1, 0, 1]模版在空间卷积(少维度的方向 padding 一个 0)得到。注意这里一定要用前向导数和后向导数的均值做模版，不能取单边模版比如[-1, 1]，因为往两个相反方向移动会得到不同的 transform 切向量 T，但是在求解二次优化问题的时候优化变量 a 当然可以取负值或正值，这样肯定会导致对切线距离的估计不准。

旋转变换的切向量 T 用原图像的每个格点的像素值减去对应图像旋转一个小角度(在实验中取为 10 度，如果太小 nearest neighbor 插值基本不 work，双线性有一定作用)后新的在该格点位置的点的像素值，该像素值用双线性插值得到。

计算一个点 x 到过另一个点 x'的变换切线的切距离的公式为:

$$\|x'-x\|^2-\frac{\left(T^T(x'-x)\right)^2}{\left\|T\right\|^2}$$

下面为分别对 N=1000, K=1 情况下 L2, L3, 三种 tangent distance 做实验得到的结果：

距离量度		Error rate
L2		0.0534
L3		0.0487
旋转	Bilinear	0.0482
水平		0.0485
垂直		0.0489

可以看出，仅基于垂直 / 水平平移变换的切线距离明显比 L2 距离效果要好，和 L3 距离效果差不多。由于旋转 transform 可以修正同类数据间细微的旋转导致的差异，可能能达到更好的效果。

在我的实验中发现，如果 N 较小，取 N=100, K=1, 基于切线距离的算法比起传统点距离量度能取得相对更大的加速。这是因为训练样本较少时，测试样本更可能找不到与自己刚好相同角度、相同位置的训练样本，需要用一些微小的变换来修正。

距离量度		Error rate
L2		0.1269
L3		0.1224
旋转	Bilinear	0.1163
	Nearest	0.1197
水平		0.1205
垂直		0.1211

下面为随手贴的一部分截图，其中 power 信息在切线距离的时候没啥用，切线距离为方便解 2 次优化问题，空间距离当然仍用 L2 距离

```
>> main
power: 3; number of points per class: 1000; K: 1: error rate: 0.048500
```

```
>> main
power: 3; number of points per class: 1000; K: 1: error rate: 0.048900
```

```
>> main
use tangent dist: rotation transform: bilinear
power: 3; number of points per class: 1000; K: 1: error rate: 0.048200
```

```
>> main
use tangent dist: rotation transform
power: 3; number of points per class: 100; K: 1: error rate: 0.116300
```

```
>> main
use tangent dist: rotation transform
power: 3; number of points per class: 100; K: 1: error rate: 0.119700
```

```
>> main
use tangent dist: horizontal movement transform
power: 3; number of points per class: 100; K: 1: error rate: 0.120500
```

```
>> main
use tangent dist: vertical movement transform
power: 3; number of points per class: 100; K: 1: error rate: 0.121100
>> main
power: 2; number of points per class: 100; K: 1: error rate: 0.126900
```