DIÁRIO DE BORDO

**1**. Uma única distribuição. 3 Variáveis.

Começamos o experimento fazendo um experimento com 3 variáveis (x0, x1, x2). Duas delas com distribuição normal e 1 totalmente aleatório. Uma das variáveis com distribuição normal (nesse caso x0) foi usada como backbone, e com ela descobrimos o ponto mais próximos.

Exemplo: Dado um ponto P1, procuramos o ponto mais próximo P2 utilizando apenas x0 como medida de proximadade. Com base nisso medíamos a distancia entre os atributos x1 e x2 dos pontos. Essa medida foi feita para todos os pontos e medidas as médias dessas distâncias.

Foi concluído que para se ter alguma conclusão as variáveis necessitavam ser normalizadas, pois se a dsitribuição das variáveis normais for em maior escala que as variáveis totalmente aleatórias as medidas de distancias ficavam “sem sentido”.

**2**. Uma única distribuição, variáveis normalizadas, N variáveis.

Em seguida todas as variáveis foram normalizadas, anromalização foi feita de forma linear utiizando os desvios padrões de cada uma das variáveis como limite superior e inferior.

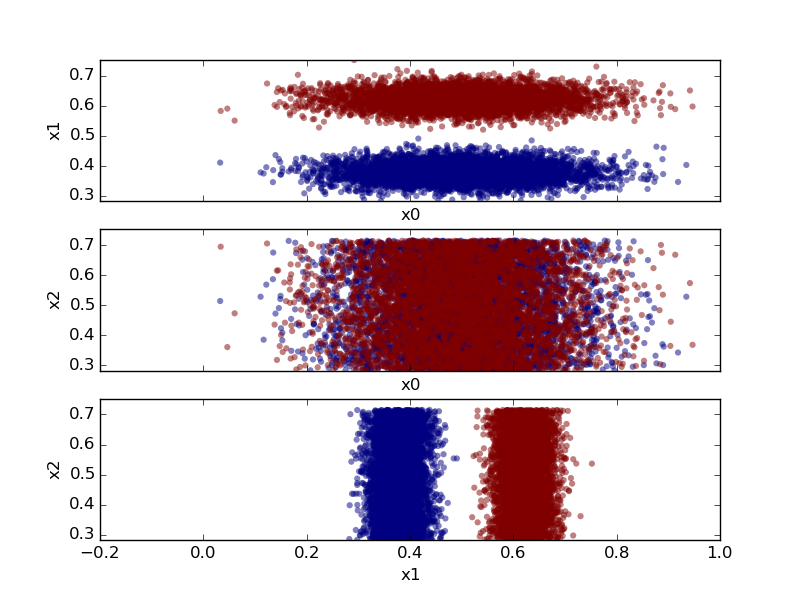
Exemplo: dada uma variável xn, ela foi normalizada usando a função f(x) = a\*x + b. Onde a e b foram calculados de forma que que f(u -[4 \* sigma]) = 0, e f(u + [4 \* sigma]) = 1. Onde u é a média de xn, e u é o desvio padrão.

A normalização provou-se útil de forma que repetindo os passos da etapa 1 **sempre** as variáveis com distribuição normal ficavam com média de distâncias menor as totalmente aleatórias. A partir daqui todos os experimentos descritos são normalizados.

Em seguida foram utilizadas mais variáveis como **backbones,** ou seja a distancia entre os pontos foi calculada utilizando mais variáveis. Como todas as variaveis foram geradas independentes uma das outras não houve diferença entre utilizar mais backbones para identificar as outras variáveis com distribuição normal.

**3**. Mais de uma distribuição. 3 variáveis. Utilizando classes ou não.

Dessa vez ao invés de uma gaussiana para as variáveis normais foram utilizadas mais de uma, como mostra a figura abaixo.



O problema surgiu que quando as gaussianas se sobrepõe no backbone (como em x0 na figura). As distancias mediasdas variáveis totalmente aleatórias ficam menor. Isso ocorre poi, para as variáveis normais, o ponto mais próximo em x0, pode estar em outra gaussiana.

Esse problema é sanado quando consideramos que cada gaussiana é uma classe diferente e calculamos a distancia somente dos pontos pertencentes a mesma classe, fazendo isso equivale a realizar varios experimentos da etapa 2 separados e depois pegar a média deles.

**Próximos passos pretendidos:**

1) Fazer experiência com dados sem ser totalmente independentes.

2) Utilizar base de dados real onde há atributos conhecidos que são importantes e outros não.

**Dúvida:**

Ainda continuo com uma dúvida. Ao que vi até aqui o que temos consiguido achar (utilizando base de dados artificiais) é qual variável tem menor aleatoriedade do que a outra. Porém não creio que isso defina ela como backbone. No exemplo da figura da Etapa 3 por exemplo, o nosso método acusaria x0 como um atributo backbone, ao passo que, em minha interpretação apenas x1 é.