1 Introduction

1.1 1NN

Takes about 19:47 minutes to run on my computer and achieves a test set accuracy of 96.61%.

1.2 FLANN

Cool idea. [TODO]

1.3 LOOCV

5-fold leave one out cross validation predicts 96.94% accuracy. Takes 1:30:45 hours to run in total.

2 Probability

2.1 Boys & girls

(Source: Minka). My neighbour has two children. Suppose I ask him whether he has any boys, and he says yes. What is the probability that one child is a girl?

$$\begin{aligned}\mathbb{P}(BG \vee GB | BB \vee BG \vee GB) &= \frac{\mathbb{P}((BG \vee GB) \wedge (BB \vee BG \vee GB))}{\mathbb{P}(BB \vee BG \vee GB)} \\ &= \frac{\mathbb{P}(BG \vee GB)}{\mathbb{P}(BB \vee BG \vee GB)} \\ &= \frac{2}{3}\end{aligned}$$

This result is sort of interesting because without the condition, the probability that one child is a girl is only $\frac{1}{2}$. It is not immediately obvious that knowing there is at least one boy would increase the probability, but this works because it cuts out the GG case, where there is not exactly one girl.

Suppose instead that I happen to see one of his children run by, and it is a boy. What is the probability that the other child is a girl?

Without loss of generality, we can assume that we saw child 1 (otherwise the events are flipped but the probabilities remain the same). Thus

$$\mathbb{P}(BG | BB \vee BG) = \frac{\mathbb{P}(BG \wedge (BB \vee BG))}{\mathbb{P}(BB \vee BG)} = \frac{\mathbb{P}(BG)}{\mathbb{P}(BB \vee BG)} = \frac{1}{2}$$

In this case, observing one child does not have any bearing on the gender of the other, whereas earlier we were given information that affected both children.

(Bonus question). The much more interesting variant of this question is when we are given that one of the children is a boy, born on a Tuesday. Now, what is the probability of both children being boys?

2.2 Legal reasoning

(Source: Peter Lee). Suppose a crime has been committed. Blood is found at the scene for which there is no innocent explanation. It is of a type which is present in 1% of the population. [Not stated, but I would guess that the defendant also has this blood type.]

The prosecutor claims: "There is a 1% chance that the defendant would have the crime blood type if he were innocent. Thus there is a 99% chance that he is guilty."

The defender claims: "The crime occurred in a city of 800,000 people. The blood type would be found in approximately 8000 people. The evidence has provided a probability of just 1 in 8000 that the defendant is guilty, and thus has no relevance."

The prosecutor's argument assumes that $\mathbb{P}(A|B) + \mathbb{P}(\neg B|A) = 1$, where A is having the crime blood type and B is being innocent. While it is true that $\mathbb{P}(A|B) = 1\%$, the rest of the statement is clearly not true in general.

The defender's argument begins with the assertion of some value for $\mathbb{P}(\neg B)$. Assuming that the criminal's blood type will match the crime scene, i.e. $\mathbb{P}(A|\neg B) = 1$, it is certainly true that $\mathbb{P}(\neg B|A) = \mathbb{P}(\neg B)/\mathbb{P}(A)$. However, this relies on the assertion that the defendant has been selected at random from the population. Since this is probably not the case, we are in fact interested in the probability $\mathbb{P}(\neg B|A \wedge D) = \mathbb{P}(\neg B|D)/\mathbb{P}(A|D)$, where D is being a defendant in the trial. If other incriminating evidence (sans blood testing) has been brought forward, it is likely that $\mathbb{P}(\neg B|D) > \mathbb{P}(\neg B)$ and $\mathbb{P}(A|D) \approx \mathbb{P}(A)$, and so $\mathbb{P}(\neg B|A \wedge D) > \mathbb{P}(\neg B|A)$. Thus the blood test has a **compounding** effect.

2.3 Variance of a sum

Suppose X and Y are random variables with means μ_X and μ_Y, respectively, and variances σ_X^2 and σ_Y^2, also respectively. Also, let $Z = X + Y$. Then

$$\begin{aligned}
\sigma_Z^2 &= \mathbb{E}[Z^2] - \mu_Z^2 \\
&= \mathbb{E}[X^2 + Y^2 + 2XY] - (\mu_X^2 + \mu_Y^2 + 2\mu_X\mu_Y) \\
&= \mathbb{E}[X^2] - \mu_X^2 + \mathbb{E}[Y^2] - \mu_Y^2 + 2(\mathbb{E}[XY] - \mu_X\mu_Y) \\
&= \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}
\end{aligned}$$

where the steps either use the definitions of variance and mean, linearity of expectation, or just plain old substitution.

2.4 Medical diagnosis

(Source: Koller).

$$\begin{aligned}\mathbb{P}(\text{disease}|\text{positive}) &= \frac{\mathbb{P}(\text{positive}|\text{disease})\mathbb{P}(\text{disease})}{\mathbb{P}(\text{positive})} \\ &= \frac{(99\%)(0.01\%)}{(99\%)(0.01\%) + (1\%)(99.99\%)} \\ &= 0.98\%\end{aligned}$$

2.5 Monty Hall problem

(Source: Mackay).

$$\begin{aligned}\mathbb{P}(1) &= \frac{1}{3} \\ \mathbb{P}(2|\neg 3) &= \frac{\mathbb{P}(\neg 3|2)\mathbb{P}(2)}{\mathbb{P}(\neg 3)} = \frac{(1)(1/3)}{2/3} = \frac{1}{2}\end{aligned}$$

To be honest, I prefer the more intuitive argument that the first pick was a choice between three options, whereas the second pick, if changed, would be a choice between two options.

2.6 Conditional independence

(Source: Koller).

$$\mathbb{P}(H = k|E_1 = e_1, E_2 = e_2) = \frac{\mathbb{P}(E_1 = e_1, E_2 = e_2|H = k)\mathbb{P}(H = k)}{\mathbb{P}(E_1 = e_1, E_2 = e_2)}$$

So, set ii. is sufficient for the calculation. If $E_1 \perp E_2|H$, then we can break down the term

$$\mathbb{P}(E_1 = e_1, E_2 = e_2|H = k) = \mathbb{P}(E_1 = e_1|H = k)\mathbb{P}(E_2 = e_2|H = k)$$

and so set i. is also sufficient. Furthermore, we can calculate the joint probability for E_1 and E_2 by marginalising over H to get

$$\begin{aligned}\mathbb{P}(E_1 = e_1, E_2 = e_2) &= \sum_{i=1}^{K} \mathbb{P}(E_1 = e_1, E_2 = e_2|H = i)\mathbb{P}(H = i) \\ &= \sum_{i=1}^{K} \mathbb{P}(E_1 = e_1|H = i)\mathbb{P}(E_2 = e_2|H = i)\mathbb{P}(H = i)\end{aligned}$$

and so all three sets actually suffice for the calculation.

2.7 Pairwise independence does not imply mutual independence

Suppose A, B, C are pairwise independent random variables. A necessary condition for mutual independence is that $\mathbb{P}(A|B,C) = \mathbb{P}(A)$, but for this to be true it would imply that

$$\mathbb{P}(A|B,C) = \frac{\mathbb{P}(B,C|A)\mathbb{P}(A)}{\mathbb{P}(B,C)} = \mathbb{P}(A)$$

and so $\mathbb{P}(B,C|A) = \mathbb{P}(B,C)$. Therefore, a counterexample would have B and C not independent given A. A simple example of this is if B and C are independent coin flips and A is whether or not they land on the same side as each other.

2.8 Conditional independence iff joint factorises

We have conditional independence $X \perp Y|Z$ iff $p(x,y|z) = p(x|z)p(y|z)$. We now show that this holds iff we can factorise the joint as $p(x,y|z) = g(x,z)h(y,z)$ for some functions g and h.

($\implies$). Suppose $p(x,y|z) = p(x|z)p(y|z)$. Let $g(x,z) = p(x|z)$ and let $h(y,z) = p(y|z)$. Done.

($\impliedby$). Suppose $p(x,y|z) = g(x,z)h(y,z)$. Then we can marginalise out y, say, as follows: $p(x|z) = \int p(x,y|z)dy = \int g(x,z)h(y,z)dy = g(x,z) \cdot \int h(y,z)dy$. Similarly for x, we have $p(y|z) = \int g(x,z)dx \cdot h(y,z)$, and so $p(x|z)p(y|z) \propto g(x,z)h(y,z) = p(x,y|z)$, since z is given (constant).

2.9 Conditional independence

(Source: Koller).

True. Suppose $(X \perp W|Z,Y) \wedge (X \perp Y|Z)$. Then

$$\begin{aligned}
p(x,w,y|z) &= p(x|z)p(w,y|x,z) && \text{(chain rule)} \\
&= p(x|z)p(w|x,y,z)p(y|x,z) && \text{(chain rule)} \\
&= p(x|z)p(w|y,z)p(y|z) && \text{(conditional independence)} \\
&= p(x|z)p(w,y|z).
\end{aligned}$$

Hence $(X \perp W|Z,Y) \wedge (X \perp Y|Z) \implies (X \perp Y,W|Z)$.

True. Unless I'm mistaken, we can prove a stronger result. Suppose $(X \perp Y|Z) \vee (X \perp Y|W)$. Then we can factorise $p(x,y|z,w)$ as either $g_z(x,z)h_z(y,z)$ or $g_w(x,w)h_w(y,w)$, depending on which factorisation is available, and so $(X \perp Y|Z) \vee (X \perp Y|W) \vee (X \perp Y|Z,W)$.

2.10 Deriving the inverse gamma density

Suppose $X \sim \text{Gamma(a, b)}$, so $p_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp^{-xb}$. If $Y = 1/X$, then

$$\begin{aligned} p_Y(y) &= p_X(y^{-1}) \left| \frac{d}{dy}(y^{-1}) \right| \\ &= \frac{b^a}{\Gamma(a)} y^{-a+1} \exp^{-b/y} y^{-2} \\ &= \frac{b^a}{\Gamma(a)} y^{-a-1} \exp^{-b/y} \end{aligned}$$

and so $Y \sim \text{Inv-Gamma}(a, b)$.

2.11 Normalisation constant for a 1D Gaussian

$$\begin{aligned} Z^2 &= \int_0^{2\pi} \int_0^{\infty} r \exp\left(-\frac{r^2}{2\sigma^2}\right) dr\, d\theta \\ &= 2\pi \left[-\sigma^2 \exp\left(-\frac{r^2}{2\sigma^2}\right) \right]_0^{\infty} \\ &= 2\pi\sigma^2. \end{aligned}$$

2.12 Expressing mutual information in terms of entropies

Recall the relevant definitions for entropy $H(X)$, conditional entropy $H(X|Y)$ and mutual information $I(X, Y)$.

$$\begin{aligned} H(X) &= -\sum_x p(x) \log p(x) \\ H(X|Y) &= -\sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(y)} \\ I(X, Y) &= \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= \sum_x \sum_y p(x, y) \left(-\log p(x) + \log \frac{p(x, y)}{p(y)} \right) \\ &= -\sum_x \sum_y p(x, y) \log p(x) + \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(y)} \\ &= -\sum_x p(x) \log p(x) - H(X|Y) \\ &= H(X) - H(X|Y). \end{aligned}$$

Similarly, $I(X, Y) = H(Y) - H(Y|X)$.

2.13 Mutual information for correlated normals

(Source: (Cover and Thomas 1991, Q9.3)). Let $\mathbf{X}$ be a random vector with a bivariate normal distribution

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}\right).$$

Then the mutual information

$$\begin{aligned} I(X_1, X_2) &= \mathbb{E}_{\mathbf{X}}\left[\log \frac{p(\mathbf{x})}{p(x_1)p(x_2)}\right] \\ &= \mathbb{E}_{\mathbf{X}}[\log p(\mathbf{x}) - \log p(x_1) - \log p(x_2)] \\ &= -H(\mathbf{X}) + H(X_1) + H(X_2) \\ &= -\frac{1}{2}\log_2\left((2\pi e)^2 \det\Sigma\right) + \log_2\left(2\pi e\sigma^2\right) \\ &= \log_2\left(\frac{\sigma^2}{\sqrt{\det\Sigma}}\right) \\ &= -\frac{1}{2}\log_2(1-\rho^2). \end{aligned}$$

Hence, when $\rho = 0$, $I(X_1, X_2) = 0$, and when $\rho^2 = 1$, $I(X_1, X_2) = \infty$. Intuitively, there is no mutual information between X_1 and X_2 when they are independent, and the opposite of this occurs when they are perfectly correlated.

2.14 A measure of correlation (normalised mutual information)

(Source: (Cover and Thomas 1991, Q2.20)). Let X and Y be discrete random variables which are identically distributed but not necessarily independent. Define

$$r = 1 - \frac{H(Y|X)}{H(X)}.$$

a. $\frac{I(X,Y)}{H(X)} = \frac{H(Y)-H(Y|X)}{H(X)} = \frac{H(X)-H(Y|X)}{H(X)} = 1 - \frac{H(Y|X)}{H(X)} = r.$

b. Entropy is non-negative, trivially, so $r \leq 1$. From the above, we can also see that if mutual information is non-negative, then $r \geq 0$.

$$\begin{aligned} I(X,Y) &= \mathbb{E}_{X,Y}\left[-\log\frac{p(x)p(y)}{p(x,y)}\right] \\ &\geq -\log \mathbb{E}_{X,Y}\left[\frac{p(x)p(y)}{p(x,y)}\right] && \text{(Jensen's)} \\ &= -\log\left(\int_{\mathbb{R}^2} p(x,y)\frac{p(x)p(y)}{p(x,y)}\,d\mathbf{x}\right) \\ &= 0. && \square \end{aligned}$$

c. $r = 0$ iff $I(X, Y) = 0$. We have equality in Jensen's iff the function is not strictly convex (but the logarithm *is* strictly convex) or when the variable inside the function is constant, i.e.

$$\frac{p(x)p(y)}{p(x,y)} = C$$

for all $x, y \in \mathbb{R}$. Due to normalisation, C must be equal to 1, and so $r = 0$ iff $p(x, y) = p(x)p(y)$, i.e. X and Y are independent.

d. $r = 1$ iff $H(Y|X) = 0$ iff $p(x, y) = p(y)\ \forall x, y \in \mathbb{R}$, i.e. X is entirely dependent (perfectly correlated with) Y, and vice versa.

2.15 MLE minimises KL divergence to the empirical distribution

Recall that the empirical distribution can be defined as

$$p_{emp}(x) = \sum_{i=1}^{N} w_i \delta_{x_i}(x)$$

where we have weights w_i for N distinct sample values x_i. The KL divergence

$$KL(p_{emp}||q) = \sum_{i=1}^{N} w_i \frac{w_i}{q(x_i)}$$

has a general minimum at 0 due to non-negativity, and this is attained when $q(x_i) = w_i\ \forall i$, which is the result of applying the MLE.

2.16 Mean, mode, variance for the beta distribution

$$\mathrm{Beta}(x|a, b) = \frac{1}{B(a, b)} x^{a-1}(1 - x)^{b-1}$$

We find the mean by repeatedly performing integration by parts.

$$\begin{aligned}
\mathbb{E}[x|a,b] &= \frac{1}{B(a,b)}\int_0^1 x^a(1-x)^{b-1}\,dx \\
&= \frac{1}{B(a,b)}\left(\frac{1}{a+1}\left[x^{a+1}(1-x)^{b-1}\right]_0^1 + \frac{b-1}{a+1}\int_0^1 x^{a+1}(1-x)^{b-2}\,dx\right) \\
&= \frac{1}{B(a,b)}\left(0 + \frac{b-1}{a+1}\int_0^1 x^{a+1}(1-x)^{b-2}\,dx\right) \\
&= \frac{1}{B(a,b)}\left(\frac{b-1}{a+1}\right)\left(0 + \frac{b-2}{a+2}\int_0^1 x^{a+2}(1-x)^{b-3}\,dx\right) \\
&= \frac{1}{B(a,b)}\left(\frac{b-1}{a+1}\right)\cdots\left(\frac{1}{a+b-1}\right)\int_0^1 x^{a+b-1}(1-x)^0\,dx \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\left(\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b)}\right)\frac{1}{a+b} \\
&= \frac{a}{a+b}.
\end{aligned}$$

The variance is calculated similarly.

$$\begin{aligned}
\mathbb{E}[x^2|a,b] &= \frac{1}{B(a,b)}\int_0^1 x^{a+1}(1-x)^{b-1}\,dx \\
&= \frac{1}{B(a,b)}\left(\frac{b-1}{a+2}\right)\cdots\left(\frac{1}{a+b}\right)\int_0^1 x^{a+b}(1-x)^0\,dx \\
&= \frac{a(a+1)}{(a+b)(a+b+1)}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}^2[x|a,b] - \mathbb{E}[x^2|a,b] &= \frac{a^2}{(a+b)^2} - \frac{a+1}{a+b+1} \\
&= \frac{a}{a+b}\left(\frac{a}{a+b} - \frac{a+1}{a+b+1}\right) \\
&= \frac{ab}{(a+b)^2(a+b+1)}.
\end{aligned}$$

We find the mode by setting the derivative to 0.

$$\begin{aligned}
0 &= \frac{d}{dx}x^{a-1}(1-x)^{b-1} \\
&= (a-1)x^{a-2}(1-x)^{b-1} - (b-1)x^{a-1}(1-x)^{b-2} \\
(b-1)x^{a-1}(1-x)^{b-2} &= (a-1)x^{a-2}(1-x)^{b-1} \\
(b-1)x &= (a-1)(1-x) \\
x &= \frac{a-1}{a+b-2}
\end{aligned}$$

2.17 Expected value of the minimum

Let $X, Y \overset{i.i.d}{\sim} U[0,1]$ and $Z = \min(X,Y)$. Normally this is done by considering the c.d.f but since we only have 2 variables here we can do it by brute force for variety.

$$\begin{aligned}
\mathbb{E}_{X,Y}[Z] &= \int_{\mathbb{R}^2} p(x,y)\min(x,y)\, d\mathbf{x} \\
&= \int_{\mathbb{R}} \int_{-\infty}^{x} p(x,y) y\, d\mathbf{x} + \int_{\mathbb{R}} \int_{x}^{\infty} p(x,y) x\, d\mathbf{x} \\
&= \int_0^1 \int_0^x y\, d\mathbf{x} + \int_0^1 \int_x^1 x\, d\mathbf{x} \\
&= \int_0^1 \frac{1}{2}x^2\, dx + \int_0^1 x(1-x) dx \\
&= \frac{1}{3}.
\end{aligned}$$