Chapter 1

Introduction

1.1 Solutions

1.1.1 KNN classifier on shuffled MNIST data

We just have to insert the following piece of code.

Listing 1.1: Part of `mnistShuffled1NNdemo`

```
... load data

%% permute columns
D = 28*28;
setSeed(0); perm = randperm(D);
Xtrain = Xtrain(:, perm);
Xtest = Xtest(:, perm);

... same as before
```

1.1.2 Approximate KNN classifiers

According to John Chia, the following code will work.

Listing 1.2: :

```
[result, ndists] = flann_search(Xtrain', Xtest', 1, ...
  struct('algorithm', 'kdtree', 'trees', 8, 'checks', 64));
errorRate = mean(ytrain(result) ~= ytest0)
```

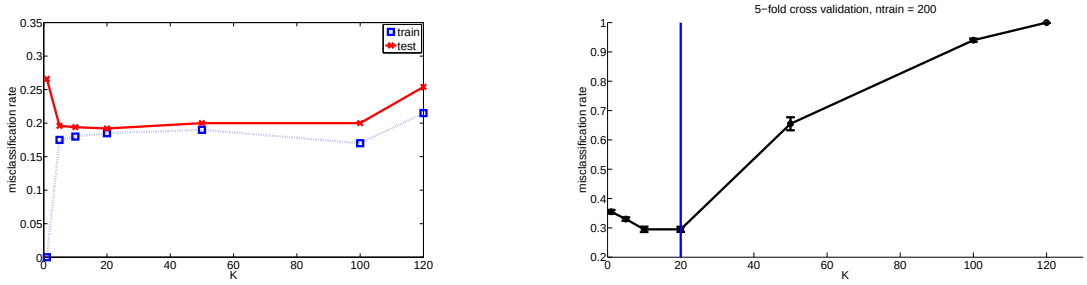
He reports the following results on MNIST with 1NN.

	ntests=1000		ntests=10,000	
	Err	Time	Err	Time
Flann	4.8%	17s	3.35%	17.2s
Vanilla	3.8%	3.68s	3.09%	28.36s

So the approximate method is somewhat faster for large test sets, but is slightly less accurate.

1.1.3 CV for KNN

See Figure 1.1(b). The CV estimate is an overestimate of the test error, but has the right shape. Note, however, that the empirical test error is only based on 500 test points. A better comparison would use a much larger test set.



(a) (b)

Figure 1.1: (a) Misclassification rate vs K in a K-nearest neighbor classifier. On the left, where K is small, the model is complex and hence we overfit. On the right, where K is large, the model is simple and we underfit. Dotted blue line: training set (size 200). Solid red line: test set (size 500). (b) 5-fold cross validation estimate of test error. Figure generated by `knnClassifyDemo`.

Chapter 2

Probability

2.1 Solutions

2.1.1 Probabilities are sensitive to the form of the question that was used to generate the answer

1. The event space is shown below, where X is one child and Y the other.

X	Y	Prob.
G	G	1/4
G	B	1/4
B	G	1/4
B	B	1/4

Let N_g be the number of girls and N_b the number of boys. We have the constraint (side information) that $N_b + N_g = 2$ and $0 \leq N_b, N_g \leq 2$. We are told $N_b \geq 1$ and are asked to compute the probability of the event $N_g = 1$ (i.e., one child is a girl). By Bayes rule we have

$$p(N_g = 1|N_b \geq 1) = \frac{p(N_b \geq 1|N_g = 1)p(N_g = 1)}{p(N_b \geq 1)} \tag{2.1}$$

$$= \frac{1 \times 1/2}{3/4} = 2/3 \tag{2.2}$$

2. Let Y be the identity of the observed child and X be the identity of the other child. We want $p(X = g|Y = b)$. By Bayes rule we have

$$p(X = g|y = b) = \frac{p(Y = b|X = g)p(X = g)}{p(Y = b)} \tag{2.3}$$

$$= \frac{(1/2) \times (1/2)}{1/2} = 1/2 \tag{2.4}$$

Tom Minka (Minka 1998) has written the following about these results:

> This seems like a paradox because it seems that in both cases we could condition on the fact that "at least one child is a boy." But that is not correct; you must condition on the event actually observed, not its logical implications. In the first case, the event was "He said yes to my question." In the second case, the event was "One child appeared in front of me." The generating distribution is different for the two events. Probabilities reflect the number of possible ways an event can happen, like the number of roads to a town. Logical implications are further down the road and may be reached in more ways, through different towns. The different number of ways changes the probability.

2.1.2 Legal reasoning

Let E be the evidence (the observed blood type), and I be the event that the defendant is innocent, and $G = \neg I$ be the event that the defendant is guilty.

1. The prosecutor is confusing $p(E|I)$ with $p(I|E)$. We are told that $p(E|I) = 0.01$ but the relevant quantity is $p(I|E)$. By Bayes rule, this is

$$p(I|E) = \frac{p(E|I)p(I)}{p(E|I)p(I) + p(E|G)p(G)} = \frac{0.01p(I)}{0.01p(I) + (1 - p(I))} \tag{2.5}$$

since $p(E|G) = 1$ and $p(G) = 1 - p(I)$. So we cannot determine $p(I|E)$ without knowing the prior probability $p(I)$. So $p(E|I) = p(I|E)$ only if $p(G) = p(I) = 0.5$, which is hardly a presumption of innocence.

To understand this more intuitively, consider the following isomorphic problem (from http://en.wikipedia.org/wiki/Prosecutor's_fallacy):

> A big bowl is filled with a large but unknown number of balls. Some of the balls are made of wood, and some of them are made of plastic. Of the wooden balls, 100 are white; out of the plastic balls, 99 are red and only 1 are white. A ball is pulled out at random, and observed to be white.

Without knowledge of the relative proportions of wooden and plastic balls, we cannot tell how likely it is that the ball is wooden. If the number of plastic balls is far larger than the number of wooden balls, for instance, then a white ball pulled from the bowl at random is far more likely to be a white plastic ball than a white wooden ball — even though white plastic balls are a minority of the whole set of plastic balls.

2. The defender is quoting $p(G|E)$ while ignoring $p(G)$. The prior odds are

$$\frac{p(G)}{p(I)} = \frac{1}{799,999} \tag{2.6}$$

The posterior odds are

$$\frac{p(G|E)}{p(I|E)} = \frac{1}{7999} \tag{2.7}$$

So the evidence has increased the odds of guilt by a factor of 1000. This is clearly relevant, although perhaps still not enough to find the suspect guilty.

2.1.3 Variance of a sum

We have

$$\begin{aligned}
\text{var}\left[X+Y\right] &= E[(X+Y)^2] - (E[X]+E[Y])^2 && (2.8)\\
&= E[X^2+Y^2+2XY] - (E[X]^2+E[Y]^2+2E[X]E[Y]) && (2.9)\\
&= E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 + 2E[XY] - 2E[X]E[Y] && (2.10)\\
&= \text{var}\left[X\right] + \text{var}\left[Y\right] + 2\text{cov}\left[X,Y\right] && (2.11)
\end{aligned}$$

If X and Y are independent, then $\text{cov}\left[X,Y\right] = 0$, so $\text{var}\left[X+Y\right] = \text{var}\left[X\right] + \text{var}\left[Y\right]$.

2.1.4 Bayes rule for medical diagnosis

Let $T = 1$ represent a positive test outcome, $T = 0$ represent a negative test outcome, $D = 1$ mean you have the disease, and $D = 0$ mean you don't have the disease. We are told

$$\begin{aligned}
P(T=1|D=1) &= 0.99 && (2.12)\\
P(T=0|D=0) &= 0.99 && (2.13)\\
P(D=1) &= 0.0001 && (2.14)
\end{aligned}$$

We are asked to compute $P(D=1|T=1)$, which we can do using Bayes' rule:

$$\begin{aligned}
P(D=1|T=1) &= \frac{P(T=1|D=1)P(D=1)}{P(T=1|D=1)P(D=1)+P(T=1|D=0)P(D=0)} && (2.15)\\
&= \frac{0.99\times 0.0001}{0.99\times 0.0001 + 0.01\times 0.9999} && (2.16)\\
&= 0.009804 && (2.17)
\end{aligned}$$

So although you are much more likely to have the disease (given that you have tested positive) than a random member of the population, you are still unlikely to have it.

2.1.5 The Monty Hall problem

Let H_i denote the hypothesis that the prize is behind door i. We make the following assumptions: the three hypotheses H_1, H_2 and H_3 are equiprobable *a priori*, i.e.,,

$$P(H_1) = P(H_2) = P(H_3) = \frac{1}{3}. \tag{2.18}$$

The datum we receive, after choosing door 1, is one of $D = 3$ and $D = 2$ (meaning door 3 or 2 is opened, respectively). We assume that these two possible outcomes have the following probabilities. If the prize is behind door 1 then the host has a free choice; in this case we assume that the host selects at random between $D = 2$ and $D = 3$. Otherwise the choice of the host is forced and the probabilities are 0 and 1.

$$\left| \begin{array}{l|l|l} P(D=2\|H_1)=\frac{1}{2} & P(D=2\|H_2)=0 & P(D=2\|H_3)=1 \\ P(D=3\|H_1)=\frac{1}{2} & P(D=3\|H_2)=1 & P(D=3\|H_3)=0 \end{array} \right| \tag{2.19}$$

Now, using Bayes theorem, we evaluate the posterior probabilities of the hypotheses:

$$P(H_i\|D=3) = \frac{P(D=3\|H_i)P(H_i)}{P(D=3)} \tag{2.20}$$

$$\left| \; P(H_1\|D=3) = \tfrac{(1/2)(1/3)}{P(D=3)} \; \right| \; P(H_2\|D=3) = \tfrac{(1)(1/3)}{P(D=3)} \; \left| \; P(H_3\|D=3) = \tfrac{(0)(1/3)}{P(D=3)} \; \right| \tag{2.21}$$

The denominator $P(D = 3)$ is $(1/2)$ because it is the normalizing constant for this posterior distribution. So

$$\left| \; P(H_1\|D=3) \;=\; \frac{1}{3} \; \right| \; P(H_2\|D=3) \;=\; \frac{2}{3} \; \left| \; P(H_3\|D=3) \;=\; 0. \; \right| \tag{2.22}$$

So the contestant should switch to door 2 in order to have the biggest chance of getting the prize.

Many people find this outcome surprising. There are two ways to make it more intuitive. One is to play the gamethirty times with a friend and keep track of the frequency with which switching gets the prize. Alternatively, you can perform a thought experiment in which the game is played with a million doors. The rules are now that the contestant chooses one door, then the game show host opens 999,998 doors in such a way as not to reveal the prize, leaving the *contestant's* selected door and *one other door* closed. The contestant may now stick or switch. Imagine the contestant confronted by a million doors, of which doors 1 and 234,598 have not been opened, door 1 having been the contestant's initial guess. Where do you think the prize is?

Another way to think about the problem is to use a directed graphical model of the form $P \rightarrow M \leftarrow F$, where P indicates the location the prize, F indicates your first choice, and M indicates which door Monty opens. Clearly P and F cause (determine) M. When we observe M, our belief about P changes because we have observed evidence about its child M.

2.1.6 Moments of a Bernoulli distribution

Mean

$$\mathbb{E}[X] = \sum_{x \in \{0,1\}} x p(x) = 0p(X=0) + 1p(X=1) = \theta \tag{2.23}$$

Variance

$$\begin{aligned} \mathrm{var}[X] &= \mathbb{E}\left[(X-\mu)^2\right] = \sum_{x \in \{0,1\}} p(x)(x-\mu)^2 && (2.24) \\ &= \theta(1-\theta)^2 + (1-\theta)(0-\theta)^2 && (2.25) \\ &= \theta(1+\theta^2-2\theta) + (1-\theta)\theta^2 && (2.26) \\ &= \theta + \theta^3 - 2\theta^2 + \theta^2 - \theta^3 && (2.27) \\ &= \theta - \theta^2 = \theta(1-\theta) && (2.28) \end{aligned}$$

Alternative proof

$$\begin{aligned} \mathbb{E}\left[X^2\right] &= 0^2 p(x=0) + 1^2 p(x=1) = \theta && (2.29) \\ \mathrm{var}[X] &= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \theta - \theta^2 = \theta(1-\theta) && (2.30) \end{aligned}$$

2.1.7 Conditional independence

1. Bayes' rule gives

$$P(H|E_1, E_2) = \frac{P(E_1, E_2|H)P(H)}{P(E_1, E_2)} \tag{2.31}$$

Thus the information in (ii) is sufficient. In fact, we don't need $P(E_1, E_2)$ because it is equal to the normalization constant (to enforce the sum to one constraint). (i) and (iii) are insufficient.

2. Now the equation simplifies to

$$P(H|E_1, E_2) = \frac{P(E_1|H)P(E_2|H)P(H)}{P(E_1, E_2)} \tag{2.32}$$

so (i) and (ii) are obviously sufficient. (iii) is also sufficient, because we can compute $P(E_1, E_2)$ using normalization.

2.1.8 Pairwise independence does not imply mutual independence

We provide two counter examples.

Let X_1 and X_2 be independent binary random variables, and $X_3 = X_1 \oplus X_2$, where $\oplus$ is the XOR operator. We have $p(X_3|X_1, X_2) \neq p(X_3)$, since X_3 can be deterministically calculated from X_1 and X_2. So the variables $\{X_1, X_2, X_3\}$ are not mutually independent. However, we also have $p(X_3|X_1) = p(X_3)$, since without X_2, no information can be provided to X_3. So $X_1 \perp X_3$ and similarly $X_2 \perp X_3$. Hence $\{X_1, X_2, X_3\}$ are pairwise independent.

Here is a different example. Let there be four balls in a bag, numbered 1 to 4. Suppose we draw one at random. Define 3 events as follows:

- X_1: ball 1 or 2 is drawn.
- X_2: ball 2 or 3 is drawn.
- X_3: ball 1 or 3 is drawn.

We have $p(X_1) = p(X_2) = p(X_3) = 0.5$. Also, $p(X_1, X_2) = p(X_2, X_3) = p(X_1, X_3) = 0.25$. Hence $p(X_1, X_2) = p(X_1)p(X_2)$, and similarly for the other pairs. Hence the events are pairwise independent. However, $p(X_1, X_2, X_3) = 0 \neq 1/8 = p(X_1)p(X_2)p(X_3)$.

2.1.9 Conditional independence iff joint factorizes

Independency $\Rightarrow$ Factorization. Let $g(x, z) = p(x|z)$ and $h(y, z) = p(y|z)$. If $X \perp Y|Z$ then

$$p(x, y|z) = p(x|z)p(y|z) = g(x, z)h(y, z) \tag{2.33}$$

Factorization $\Rightarrow$ Independency. If $p(x, y|z) = g(x, z)h(y, z)$ then

$$1 = \sum_{x,y} p(x, y|z) = \sum_{x,y} g(x, z)h(y, z) = \sum_x g(x, z) \sum_y h(y, z) \tag{2.34}$$

$$p(x|z) = \sum_y p(x, y|z) = \sum_y g(x, z)h(y, z) = g(x, z) \sum_y h(y, z) \tag{2.35}$$

$$p(y|z) = \sum_x p(x, y|z) = \sum_x g(x, z)h(y, z) = h(y, z) \sum_x g(x, z) \tag{2.36}$$

$$p(x|z)p(y|z) = g(x, z)h(y, z) \sum_x g(x, z) \sum_y g(y, z) \tag{2.37}$$

$$= g(x, z)h(y, z) = p(x, y|z) \tag{2.38}$$

2.1.10 Conditional independence

1. True, since

$$(X \perp W|Z, Y) \Rightarrow p(X|W, Z, Y) = p(X|Z, Y) \tag{2.39}$$

$$(X \perp Y|Z) \Rightarrow p(X|Z, Y) = p(X|Z) \tag{2.40}$$

$$\Rightarrow p(X|W, Z, Y) = p(X|Z) \tag{2.41}$$

$$\Rightarrow (X \perp Y, W|Z) \tag{2.42}$$

2. False. Consider the DAG in Figure 2.1. It encodes that $(X \perp Y|Z)$ and $(X \perp Y|W)$ but not $(X \perp Y|Z, W)$.