Bayesian analysis of a threshold classifier

Thomas P. Minka

May 2, 2002

Abstract

For a threshold classifier we can do an exact Bayesian analysis, which reveals the exact relationship between the margin and the generalization error.

1 Introduction

For discriminating two classes, a threshold classifier assigns a class y to measurement x according to:

$$y = \mathrm{sign}(x - t) \tag{1}$$

for some threshold parameter t. Given a training set of labeled measurements $D = \{(x_1, y_1), ..., (x_n, y_n)\}$, we want to classify future data, give a measure of confidence in our classification, and estimate our generalization error.

2 Zero noise

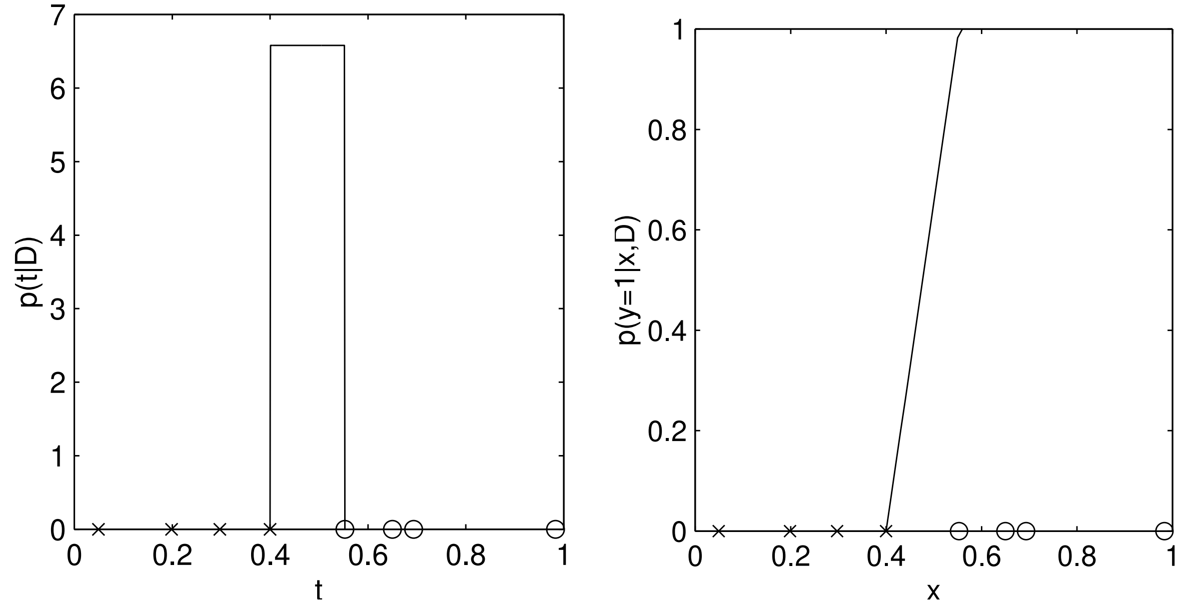
Zero noise means that the classes really are separable by a threshold. Given x and t, y is known deterministically. As a statistical model, this means that the probability that $y = 1$ is either 0 or 1, according to a step function:

$$\begin{aligned} p(y = 1|x, t) &= \Theta(x - t) && (2)\\ \Theta(x) &= \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} && (3)\\ p(y = -1|x, t) &= 1 - p(y = 1|x, t) && (4) \end{aligned}$$

Given the training set D, the likelihood for t is

$$p(D|t) = \prod_i p(y_i|x_i, t) = \begin{cases} 1 & \text{if } t \text{ classifies the data perfectly} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The maximum-likelihood estimate for t is not unique—any threshold which classifies the training data correctly is an MLE.



(a) (b)

Figure 1: (a) The posterior for t in the zero-noise model. (b) The predictive density for y, which is also the CDF for t.

To make things concrete, suppose $x \in [0, 1]$ so that $t \in [0, 1]$. For the Bayesian analysis we choose a uniform (shift-invariant) prior:

$$p(t) \sim U(0, 1) \tag{6}$$

The posterior for t is therefore proportional to the likelihood:

$$p(t|D) \propto \prod_i p(y_i|x_i, t) \tag{7}$$

As shown in figure 1(a), it is uniform between x_l, the largest measurement labeled $y = -1$, and x_h, the smallest measurement labeled $y = 1$. In machine learning parlance, this interval is the *version space.*

To classify a test point x, we can use the predictive density for y given x and the data, with the true threshold integrated out:

$$\begin{aligned} p(y = 1|x, D) &= \int_0^1 p(y = 1|x, t)p(t|D)dt & (8) \\ &= \int_0^1 \Theta(x - t)p(t|D)dt & (9) \\ &= \int_0^x p(t|D)dt = Pr(t \leq x|D) & (10) \end{aligned}$$

This is simply the cumulative density of t given the training set (figure 1(b)). To minimize the expected test error rate, we should choose $y = 1$ if $p(y = 1|x, D) > 1/2$ and $y = -1$ otherwise. Because the cumulative density is monotonic, this is equivalent to a threshold classifier where t is the point at which $p(y = 1|x = t, D) = 1/2$. This happens at $\hat{t} = (x_h + x_l)/2$, the midpoint of the version space.

The same optimal $\hat{t}$ can be derived from decision theory. For simplicity, assume the distribution of test measurements is uniform: $p(x) = 1$. If we classify according to $\hat{t}$ but the true threshold is t, then our expected test error is $|\hat{t} - t|$, because every measurement between t and $\hat{t}$ is misclassified while all other measurements are correctly classified. Using our posterior for t, the expected generalization error of $\hat{t}$ is therefore

$$G(\hat{t}) = \int_0^1 |\hat{t} - t| \; p(t|D) dt \tag{11}$$

which we recognize as the expected loss under an absolute-difference loss function. The optimal $\hat{t}$ must be the *median* of the posterior $p(t|D)$, which is the midpoint of version space. If $p(x)$ is not uniform, then the generalization error is different but the optimal $\hat{t}$ is still the median.

Because of the simple form of the posterior, we can compute the generalization error of any estimate explicitly:

$$\begin{aligned} G(\hat{t}) &= \int_{x_l}^{\hat{t}} \frac{\hat{t} - t}{x_h - x_l} dt + \int_{\hat{t}}^{x_h} \frac{t - \hat{t}}{x_h - x_l} dt & (12) \\ &= \frac{(\hat{t} - x_l)^2 + (x_h - \hat{t})^2}{2(x_h - x_l)} \qquad \text{if } \hat{t} \in [x_l, x_h] & (13) \\ &= \frac{\left(\hat{t} - \frac{x_h + x_l}{2}\right)^2}{x_h - x_l} + \frac{x_h - x_l}{4} \qquad \text{if } \hat{t} \in [x_l, x_h] & (14) \end{aligned}$$

If we define the *dataset margin* M and *margin of* t $M(t)$:

$$\begin{aligned} M &= x_h - x_l & (15) \\ M(t) &= \min_i |t - x_i| & (16) \\ G(\hat{t}) &= \frac{M(t)^2 + (M - M(t))^2}{2M} \qquad \text{if } \hat{t} \in [x_l, x_h] & (17) \end{aligned}$$

then we see that the generalization error is lower when $M(t)$ is higher, as long as M is fixed. This justifies the statement that a classifier with bigger margin is better. However, note that the generalization error of the best classifier is

$$G\left(\frac{x_h + x_l}{2}\right) = \frac{M}{4} \tag{18}$$

which is increasing in the dataset margin. Across different datasets, small dataset margin leads to better performance, because the threshold is better determined. So we have to be careful in using $M(t)$ to assess generalization error. For example, if we do feature selection then we are altering the dataset and making $M(t)$ bigger is not necessarily better.

3 Logistic error model

Now consider the case where the classes are not separable by a threshold, due to the fact that the labeling process is inherently ambiguous. Ambiguity usually happens because we have a limited set of measurements. Ambiguity means that knowing x does not determine y completely; it only affects the probability that $y = 1$ or -1. Consider the following statistical model:

$$p(y = 1|x, t, a) = \sigma(a(x - t)) \tag{19}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{20}$$

It can also be written in a symmetric way for $y = \pm 1$:

$$p(y|x, t, a) = \sigma(ay(x - t)) \tag{21}$$

In this model, the probability that $y = 1$ increases smoothly with x. The natural boundary between the classes is at t, where our uncertainty about y is highest. The scale parameter a controls the amount of ambiguity: for large a, $\sigma(ax) \approx \Theta(x)$ and we recover the zero noise model. For small a, $\sigma(ax) \approx 1/2$ and x gives essentially no information about y.

Assuming independent noise on each label, the likelihood is

$$p(D|t, a) = \prod_i p(y_i|x_i, t, a) \tag{22}$$

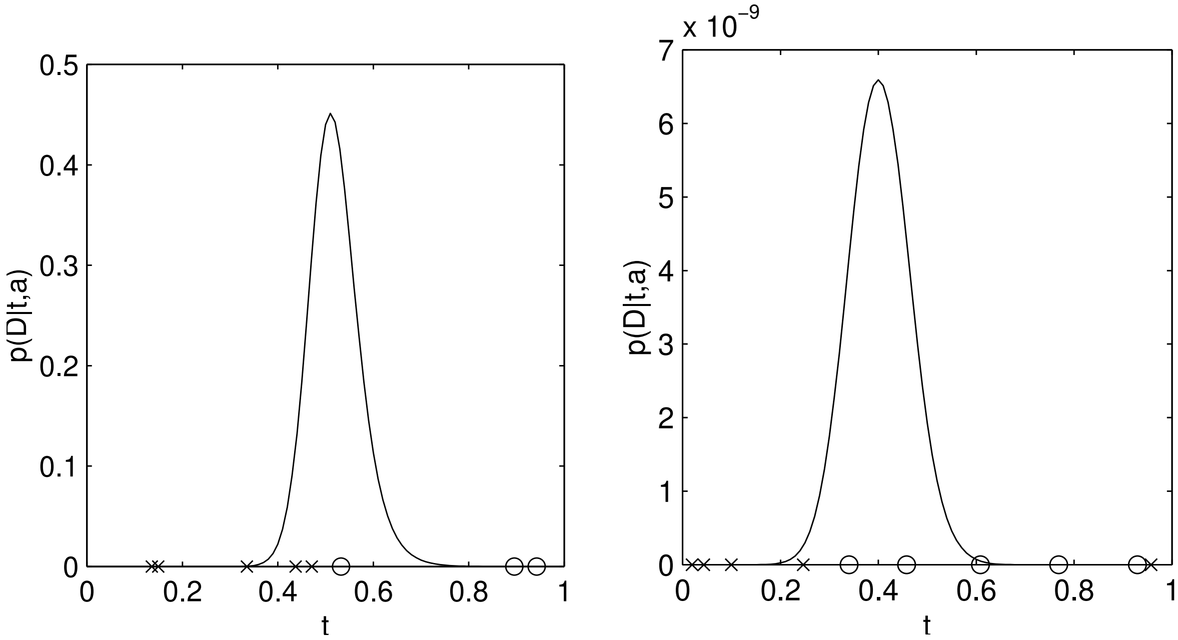
Figure 2 shows the basic shape. Given a, the likelihood for t is always unimodal and asymmetric, with falloff depending on the density of the measurements in each class. There is always a unique MLE for t, whether a is known or not. However, the MLE does not necessarily minimize training set error (see the next section). This happens especially when there is an outlier (figure 2(b)). Since the model only allows errors near the boundary, the boundary must move to accomodate the outlier and incur an additional training set error.

If the training data is separable by a threshold, then the MLE for a is infinite, and the MLE for t will limit to a value $\hat{t}$. We can deduce this value as follows. When a is large, $e^{-ay(x-t)}$ will be small, and when b is small, $\log(1 + b) \approx b$. Therefore the log-likelihood for large a is

$$\log p(D|t, a) = -\sum_i \log(1 + \exp(-ay_i(x_i - t))) \approx -\sum_i \exp(-ay_i(x_i - t)) \tag{23}$$

This summation is dominated by the data points with minimal value of $y_i(x_i - t)$, i.e. the points closest to the boundary. The optimal t will maximize its distance from these points, while classifying them correctly. Therefore $\hat{t} = (x_h + x_l)/2$, the midpoint of version space as in section 2. In that section, we arrived at the midpoint via a Bayesian argument; here we obtain it from maximum likelihood alone.

A Bayesian analysis of the logistic regression model is not possible analytically so we skip to the next model.



(a) (b)

Figure 2: (a) The likelihood function for t in the logistic model, with $a = 30$. (b) The effect of an outlier. The MLE makes two errors on the training set, while $t = 0.3$ would have made one error.

4 Uniform error model

An alternative to label noise caused by ambiguity is for the labels to be uncertain with a fixed probability e:

$$p(y = 1|x, t, e) \quad = \quad e + (1 - 2e)\Theta(x - t) \tag{24}$$

This model is the most robust in the sense that no matter how the data were really generated or what the decision boundary is, you will still get reasonable results. Any descrepancy between the model and reality will be absorbed in the error rate e. This error rate is called the "intrinsic error rate" because it is unavoidable regardless of how much training data you have.

Assuming independent noise on each label, the likelihood is

$$p(D|t, e) = \prod_i p(y_i|x_i, t, e) = e^{r(t)}(1 - e)^{n-r(t)} \tag{25}$$

where $r(t)$ is the number of training errors of threshold t. The likelihood can be multimodal, and the maximum-likelihood estimate for t is not unique—any threshold with minimal training set error is an MLE. The function $r(t)$ has at most n different values. Define r_i to be $r(t)$ for $t \in [x_i, x_{i+1}]$ (for r_0 use $t \in [0, x_1]$ and for r_n use $t \in [x_n, 1]$). The likelihood is composed of boxes of width $x_{i+1} - x_i$ and height $e^{r_i}(1 - e)^{n-r_i}$. Define A_i to be the area of the ith box:

$$A_i(e) = (x_{i+1} - x_i)e^{r_i}(1 - e)^{n-r_i} \tag{26}$$

Let B_i be the normalized area of box i: $B_i(e) = \frac{A_i(e)}{\sum_{i=0}^{n} A_i(e)}$.

With a uniform prior on t, the posterior is proportional to the likelihood:

$$p(t|D, e) \quad = \quad \frac{e^{r(t)}(1 - e)^{n-r(t)}}{p(D|e)} \tag{27}$$

$$p(D|e) \quad = \quad \sum_{i=0}^{n} A_i(e) \tag{28}$$

$$p(t|D, e) \quad = \quad \frac{B_i(e)}{x_{i+1} - x_i} \qquad \text{if } t \in [x_i, x_{i+1}] \tag{29}$$

Figure 3 shows the basic shape of the posterior. When e is known, the predictive density is related to the CDF for t as before:

$$p(y = 1|x, D, e) \quad = \quad \int_0^1 (e + (1 - 2e)\Theta(x - t))p(t|D, e)dt \tag{30}$$

$$= \quad e + (1 - 2e)Pr(t \leq x|D, e) \tag{31}$$

$$Pr(t \leq x|D, e) \quad = \quad \sum_j B_j(e)\Theta(x - x_j) + \frac{x - x_{i+1}}{x_{i+1} - x_i}B_i(e) \qquad t \in [x_i, x_{i+1}] \tag{32}$$

The Bayes-optimal threshold satisfies $p(y = 1|x = t, D, e) = 1/2$, which does not necessarily minimize training set error (figure 3(b)). So with this model we have a significant difference between Bayes and the MLE.

The Bayes-optimal threshold can also be derived from decision theory. Assume the distribution of test measurements is uniform: $p(x) = 1$. If we classify according to $\hat{t}$ but the true threshold is t, then our expected test error is $e + (1-2e)|\hat{t} - t|$. The expected generalization error is

$$G(\hat{t}|e) = \int_0^1 (e + (1-2e)|\hat{t} - t|)\ p(t|D,e)dt = e + (1-2e)\int_0^1 |\hat{t} - t|\ p(t|D,e)dt \tag{33}$$

which means the optimal $\hat{t}$ is the median of the posterior distribution for t. The same result holds when $p(x)$ is not uniform.

The generalization error can be computed by breaking the range of integration into the separate intervals $(0, x_1), (x_1, x_2), ..., (x_n, 1)$. Define the function

$$W(\hat{t}; a, b) = \frac{1}{b-a}\int_a^b |\hat{t} - t|\ dt = \begin{cases} \dfrac{(\hat{t} - \frac{a+b}{2})^2}{b-a} + \dfrac{b-a}{4} & \text{if } \hat{t} \in [a,b] \\ |\hat{t} - \frac{a+b}{2}| & \text{otherwise} \end{cases} \tag{34}$$

Then

$$\begin{aligned} \int_0^1 |\hat{t} - t|\ p(t|D,e)dt &= \int_0^{x_1} \frac{|\hat{t} - t|B_0(e)}{x_1}dt + \int_{x_1}^{x_2} \frac{|\hat{t} - t|B_1(e)}{x_2 - x_1}dt + \cdots + \int_{x_n}^1 \frac{|\hat{t} - t|B_n(e)}{1 - x_n}dt & (35) \\ &= W(\hat{t}; 0, x_1)B_0(e) + W(\hat{t}; x_1, x_2)B_1(e) + \cdots + W(\hat{t}; x_n, 1)B_n(e) & (36) \end{aligned}$$

Figure 4 plots $G(\hat{t})$ for the previous two datasets.

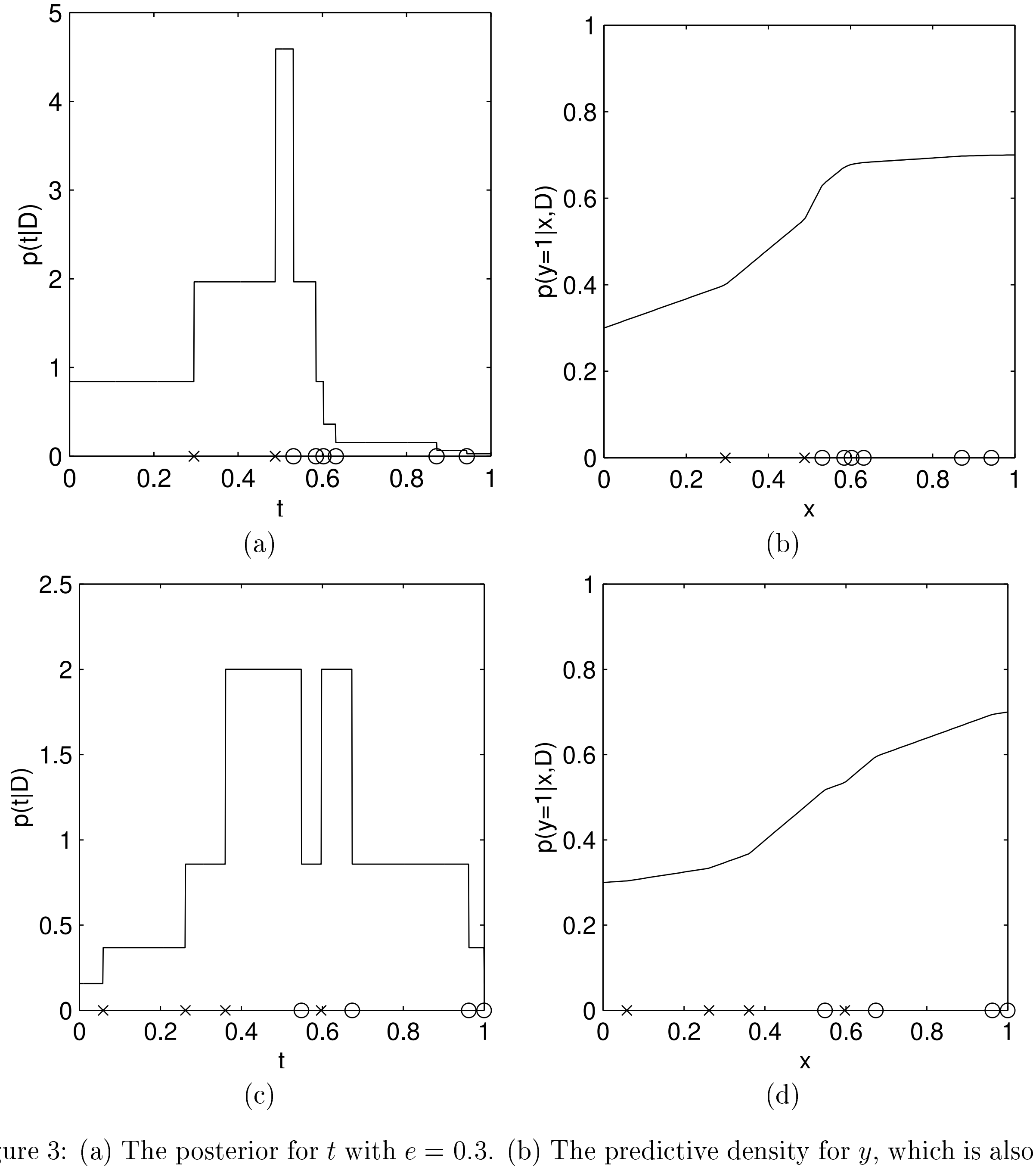


Figure 3: (a) The posterior for t with $e = 0.3$. (b) The predictive density for y, which is also the CDF for t. Note that in this example the optimal $t = 0.4$, which makes one training set error. (c,d) A multimodal posterior and its predictive density.

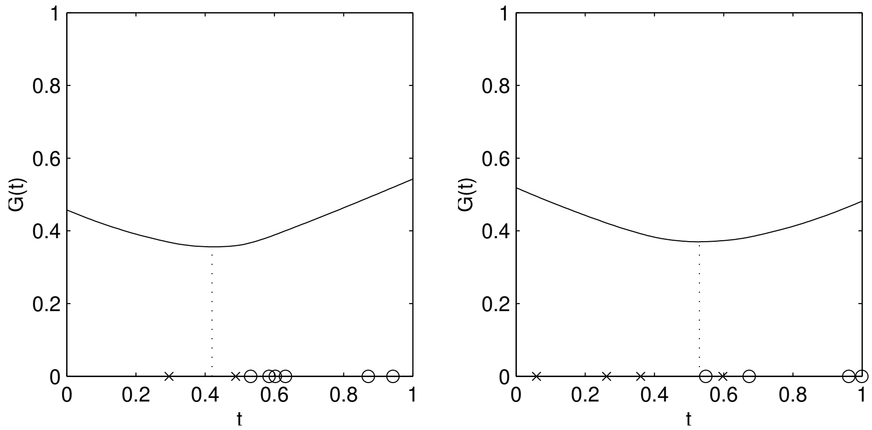


Figure 4: Expected generalization error for each threshold, with a dotted line through the minimum. The intrinsic error rate is known to be $e = 0.3$ as in the previous figure.

4.1 Unknown e

The Bayesian analysis can also be done when e is unknown. The calculations are the same but with different box areas. Let e have a Beta prior distribution:

$$p(e) \sim \mathcal{B}(\alpha_1, \alpha_2) \tag{37}$$

Then the marginal likelihood for t is

$$p(D|t) = \int_0^1 p(D|t,e)p(e)de = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(n+\alpha_1+\alpha_2)} \frac{\Gamma(r(t)+\alpha_1)}{\Gamma(\alpha_1)} \frac{\Gamma(n-r(t)+\alpha_2)}{\Gamma(\alpha_2)} \tag{38}$$

We can define the new box areas:

$$A_i = (x_{i+1}-x_i)\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(n+\alpha_1+\alpha_2)} \frac{\Gamma(r_i+\alpha_1)}{\Gamma(\alpha_1)} \frac{\Gamma(n-r_i+\alpha_2)}{\Gamma(\alpha_2)} \tag{39}$$

The marginal posterior for t is

$$p(t|D) = \frac{B_i}{x_{i+1}-x_i} \qquad \text{if } t \in [x_i, x_{i+1}] \tag{40}$$

$$p(D) = \sum_{i=0}^{n} A_i \tag{41}$$

The marginal posterior for e is

$$p(e|D) = \frac{p(D|e)p(e)}{p(D)} \tag{42}$$

where $p(D|e)$ is defined by (28). The expected value of e is

$$E[e|D] = \sum_{i=0}^{n} \frac{r_i+\alpha_1}{n+\alpha_1+\alpha_2} B_i \tag{43}$$

Figure 5 plots these posteriors for the previous two datasets (cf figure 3).

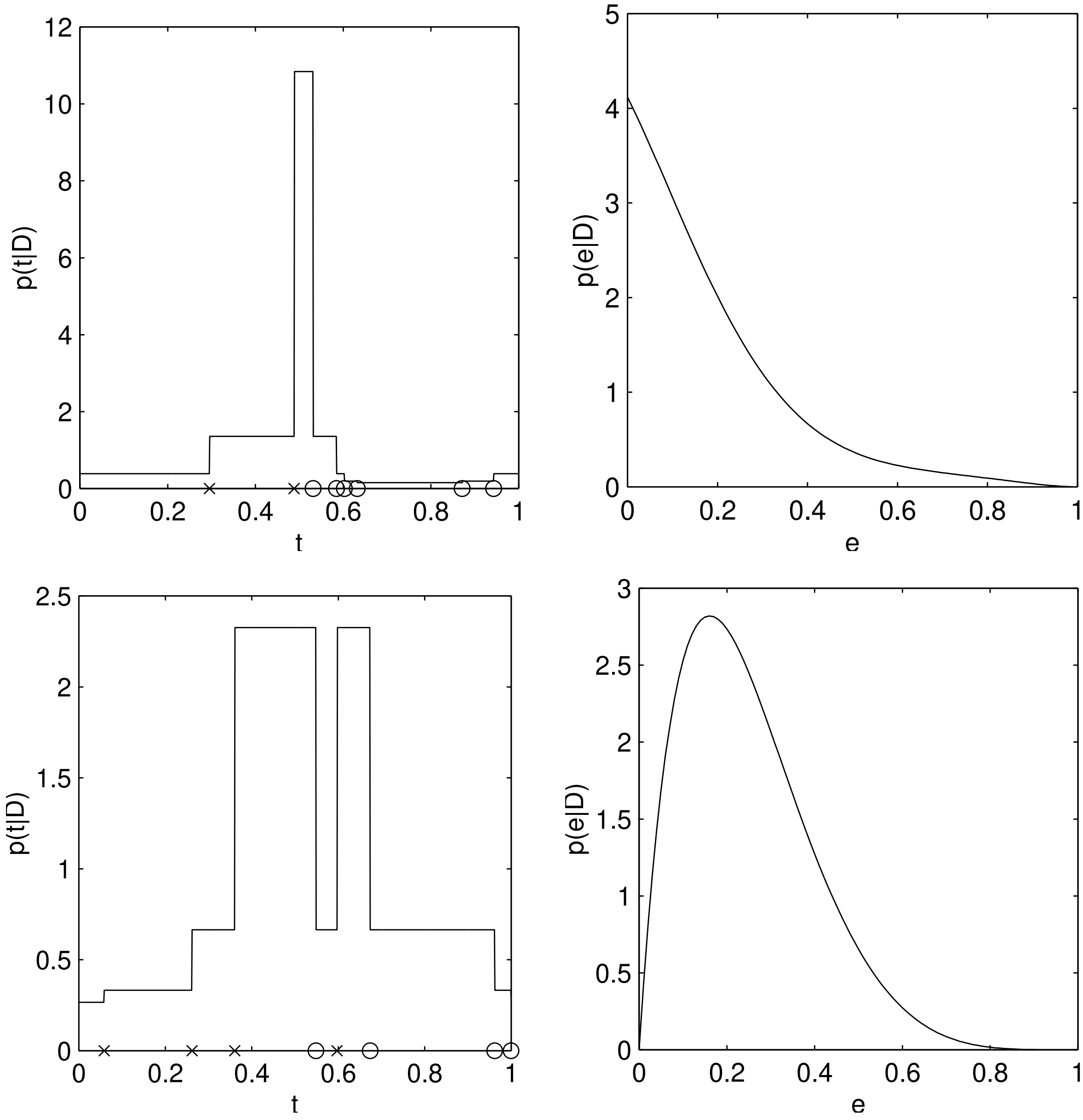


Figure 5: The posterior for t and for e when e has a uniform prior ($\alpha_1 = \alpha_2 = 1$).

For the predictive density, note that

$$\int_0^1 (1-2e)p(D|t,e)p(e)de \;=\; \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(n+\alpha_1+\alpha_2)}\frac{\Gamma(r(t)+\alpha_1)}{\Gamma(\alpha_1)}\frac{\Gamma(n-r(t)+\alpha_2)}{\Gamma(\alpha_2)}\left(1-2\frac{r(t)+\alpha_1}{n+\alpha_1+\alpha_2}\right) \tag{44}$$

Define the modified box areas

$$B'_i \;=\; B_i\left(1-2\frac{r_i+\alpha_1}{n+\alpha_1+\alpha_2}\right) \tag{45}$$

Then

$$\begin{aligned} p(y=1|x,D) &= \int_0^1\int_0^1 (e+(1-2e)\Theta(x-t))p(t,e|D)dedt & (46)\\ &= E[e|D] + \int_0^1 \frac{\Theta(x-t)p(t)}{p(D)}\left(\int_0^1 (1-2e)p(D|t,e)p(e)de\right)dt & (47)\\ &= E[e|D] + \sum_j B'_j\Theta(x-x_j) + \frac{x-x_{i+1}}{x_{i+1}-x_i}B'_i \qquad x\in[x_i,x_{i+1}] & (48) \end{aligned}$$

The expected generalization error is

$$\begin{aligned} G(\hat{t}) &= \int_0^1\int_0^1 (e+(1-2e)|\hat{t}-t|)\; p(t,e|D)dtde & (49)\\ &= E[e|D] + \int_0^1 \frac{|\hat{t}-t|p(t)}{p(D)}\left(\int_0^1 (1-2e)p(D|t,e)p(e)de\right)dt & (50)\\ &= E[e|D] + \sum_i W(\hat{t};x_i,x_{i+1})B'_i & (51) \end{aligned}$$

Figure 6 plots the predictive density and generalization functions for the same two datasets.

The generalization error is the sum of two quantities: the intrinsic error e and the estimation error of t. These two act in opposite ways. For example, if the training data is separable then the estimation error increases with the dataset margin (section 2). However, the expected intrinsic error $E[e|D]$ goes down with a bigger dataset margin, since it is more likely that the intrinsic error is zero. Thus there is an "ideal" margin value which balances the two effects and induces the lowest generalization error. Figure 7 shows the outcome of a Monte Carlo experiment using 500 random datasets of size $n = 8$. The datasets were drawn according to $p(x) = 1$ and $y = \text{sign}(x - 0.5)$. Larger margin generally makes $E[e|D]$ smaller, though there are exceptions since $E[e|D]$ depends on more than just the margin. The generalization error curve clearly reveals the two effects. It also demonstrates that the margin alone is not sufficient to predict generalization. Note that with $n = 8$ the minimum possible value of $E[e|D]$ is 0.1. Increasing n lowers the expected errors and the margins but does not change the overall shape of the curves. Finding a simple function of the data which does correlate well with $G(\hat{t})$ is an important research problem.

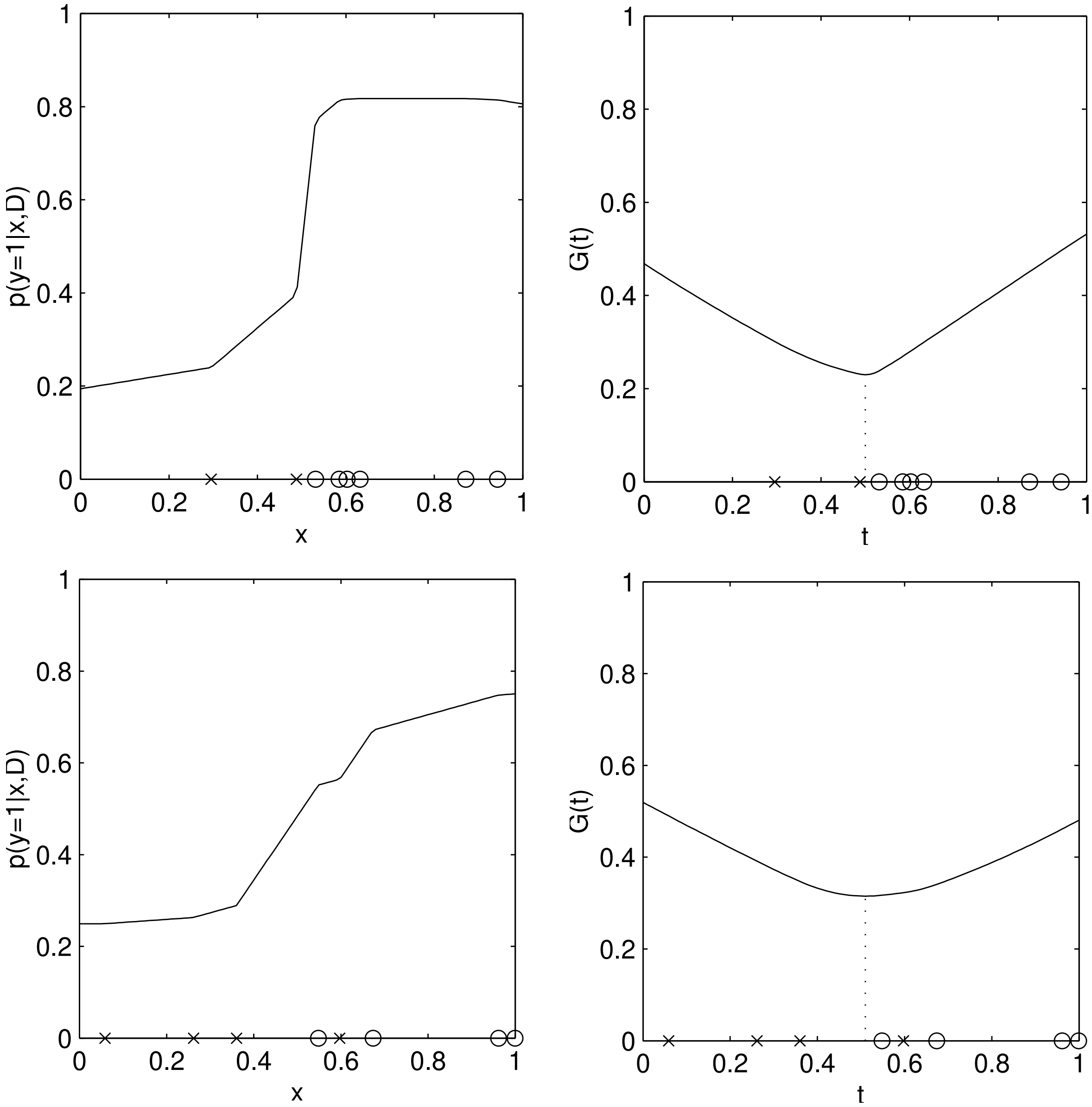


Figure 6: The predictive density $p(y = 1|x, D)$ and generalization function $G(\hat{t})$ when e has a uniform prior density ($\alpha_1 = \alpha_2 = 1$).

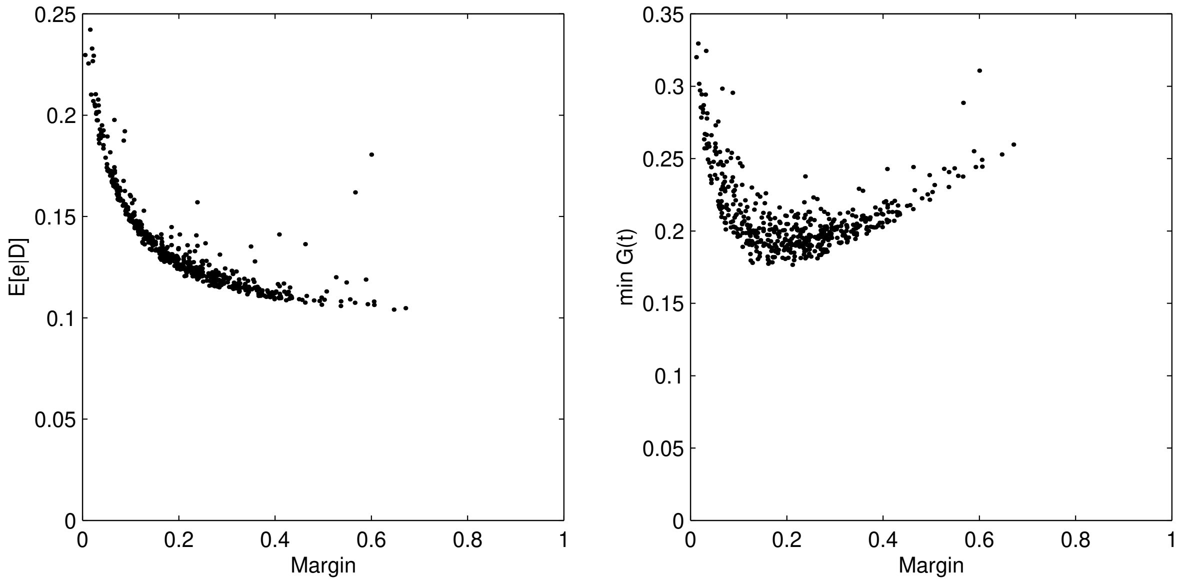


Figure 7: The relationship between the dataset margin, the expected intrinsic error e, and the expected generalization error over 500 datasets.