Bayesian Kernel Methods

THE AUSTRALIAN NATIONAL UNIVERSITY

Unit 1: Bayes Rule, Approximate Inference, Hyperparameters

Unit 2: Gaussian Processes, Covariance Function, Kernel

Unit 3: GP: Regression

Unit 4: GP: Classification

Unit 5: Implementation: Laplace Approximation, Low Rank Methods

Unit 6: Implementation: Low Rank Methods, Bayes Committee Machine

Unit 7: Relevance Vector Machine: Priors on Coefficients

Unit 8: Relevance Vector Machine: Efficient Optimization and Extensions

http://mlg.anu.edu.au/∼smola/summer2002/

Overview of Unit 4: GP Classification

THE AUSTRALIAN NATIONAL UNIVERSITY

01: Estimating Probabilities
02: Logistic Regression
03: Multiclass Logistic Regression
04: Probit Model
05: Label Noise
06: Discriminant Analysis
07: MAP Approximation
08: Optimization Problems (Why Logit is good for you)
09: Laplace Approximation and Error Bars
10: Examples
11: Hyperparameters
12: Soft Margin Loss
13: How to fix it
14: Platt's Trick
15: Why all is well (Proof by Graph)
16: Scaling Problems

Estimating Probabilities

THE AUSTRALIAN NATIONAL UNIVERSITY

Classification Problem

Unlike in regression we have $y_i \in \mathcal{Y}$ with $|\mathcal{Y}| \in \mathbb{N}$, in other words, we have only a finite number of possible outcomes. Again, the goal is to estimate $p(y|x_i)$.

Special Case

Consider the binary classification problem where $\mathcal{Y} = \{\pm 1\}$.

Problem

It is easy to build estimators generating unconstrained functions $f(x)$, yet we need some tricks to make sure that p is normalized, i.e., $\sum_u p(y|x) = 1$.

Solution

We use a link function $l(y, f(x), x)$ connecting a real valued function f and $p(y|x, f) = l(y, f(x), x)$.

Logistic Regression

Basic Idea

For classification purposes we are mainly interested in the ratio between $p(y = 1|x)$ and $p(y = -1|x)$, since this tells us the Bayes optimal classifier (i.e., the classifier with minimal error).

Making the Problem Symmetric

Estimating $\frac{p(y=1|x)}{p(y=-1|x)}$ would help us find a classifier, but it isn't symmetric with respect to y. So we attempt to find f with

$$f(x) = \log \frac{p(y = 1|x)}{p(y = -1|x)} \Rightarrow p(y = 1|x) = \frac{1}{1 + \exp(-f(x))}.$$

Likewise $p(y = -1|x) = \frac{1}{1+\exp(f(x))}$,

Likelihood

For the likelihood we obtain

$$p(Y|X, f) = \prod_{i=1}^{m} \frac{1}{1 + \exp(-y_i f(x_i))} \Rightarrow -\log p(Y|X, f) = \sum_{i=1}^{m} \log(1+\exp(-y_i f(x_i))).$$

Multiclass Logistic Regression

Observation

We may write $p(y|x, f(x))$ as follows

$$
\begin{aligned}
p(y = 1|x, f(x)) &= \frac{\exp(\frac{1}{2}f(x))}{\exp(\frac{1}{2}f(x)) + \exp(-\frac{1}{2}f(x))} \\
p(y = -1|x, f(x)) &= \frac{\exp(-\frac{1}{2}f(x))}{\exp(\frac{1}{2}f(x)) + \exp(-\frac{1}{2}f(x))}
\end{aligned}
$$

Idea

For more than two classes, estimate one function $f_j(x)$ per class and compute probabilities $p(y_j|x, f)$ via

$$p(y_j|x, f) = \frac{\exp(f_j(x))}{\sum_{i=1}^{N} \exp(f_i(x))}$$

Posterior

$$p(f|X, Y) \propto \prod_{i=1}^{m} \frac{\exp(f_{y_i}(x_i))}{\sum_{i=1}^{N} \exp(f_i(x_i))} \prod_{j=1}^{N} p(f_j)$$

Probit Model

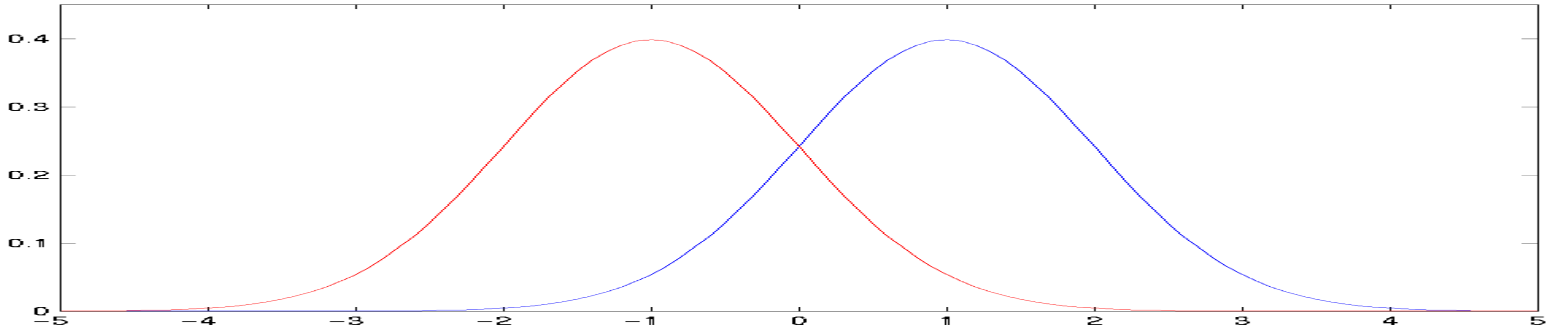
THE AUSTRALIAN NATIONAL UNIVERSITY

Basic Idea

We may assume that y is given by the sign of f, but corrupted by Gaussian noise; thus, $y = \operatorname{sgn}(f(x) + \xi)$ where $\xi \sim \mathcal{N}(0, \sigma)$. In this case, we have

$$\begin{aligned} p(y|f(x)) &= \int \frac{\operatorname{sgn}(yf(x) + \xi) + 1}{2} p(\xi) d\xi \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-yf(x)}^{\infty} \exp\left(-\frac{\xi^2}{2\sigma^2}\right) d\xi = \Phi\left(\frac{yf(x)}{\sigma}\right). \end{aligned}$$

Here Φ is the distribution function of the normal distribution.



Label Noise

THE AUSTRALIAN NATIONAL UNIVERSITY

Basic Idea

We want to perform classification in the presence of random label noise (in addition to the noise model $p_0(y|t)$ discussed previously).

Here, a label is randomly *assigned* to observations with probability 2η (note that this is the same as randomly *flipping* with probability η). We then write

$$p(y|f(x)) = \eta + (1 - 2\eta)p_0(y|f(x)).$$

Consequence

The influence of $p_0(y|f(x))$ on the posterior is descreased, hence η has a "regularizing" effect on the estimate.

Discriminant Analysis

Basic Idea

Assume that the classes to be separated (we assume $N = 2$ for simplicity) correspond to **Normal distributions** in some space, and that $f(x)$ are **projections** from this space onto a line.

Result

Projections on a real line yield normal distributions. Hence we can model the probability $p(y|x, f(x))$ by

$$p(y|x, f(x)) \propto \exp\left(-\frac{1}{2}(y - f(x))^2\right).$$

Algorithmic Result

This is essentially **regression on the labels**, which can be done very cheaply.

Problem: often the assumption of a normal distribution is not so well satisfied.

MAP Approximation

THE AUSTRALIAN NATIONAL UNIVERSITY

Log-Posterior

Instead of integrating over $p(f|X,Y)$ we minimize the negative log-posterior. To make matters simpler, we reparameterize $f = K\alpha$.

$$-\log p(f|X,Y) = \sum_{i=1}^{m} -\log l(y_i, x_i, [K\alpha]_i) + \frac{1}{2}\alpha^\top K\alpha.$$

Practical Issues

- Convex loss functions lead to optimization problems with a **global minimum**. Proof: assume two (local) minima at, say $\mathbf{t}_1, \mathbf{t}_2$, then for all arguments $\lambda\mathbf{t}_1 + (1-\lambda)\mathbf{t}_2$ the values will be less or equal to the linear interpolation. This, however, is a contradiction.
- Choice of link function determines convexity of the optimization problem.
- Morale of the story: choose link function according to data **and** numerical considerations.

Examples

THE AUSTRALIAN NATIONAL UNIVERSITY

Penalized Logistic Regression

We use the logistic link function, which leads to the following minimization problem:

$$\text{minimize} \sum_{i=1}^{m} \log\left(1 + \exp\left(-y_i \sum_{j=1}^{m} k(x_i, x_j)\alpha_j\right)\right) + \frac{1}{2}\alpha^\top K\alpha$$

where $f = K\alpha$

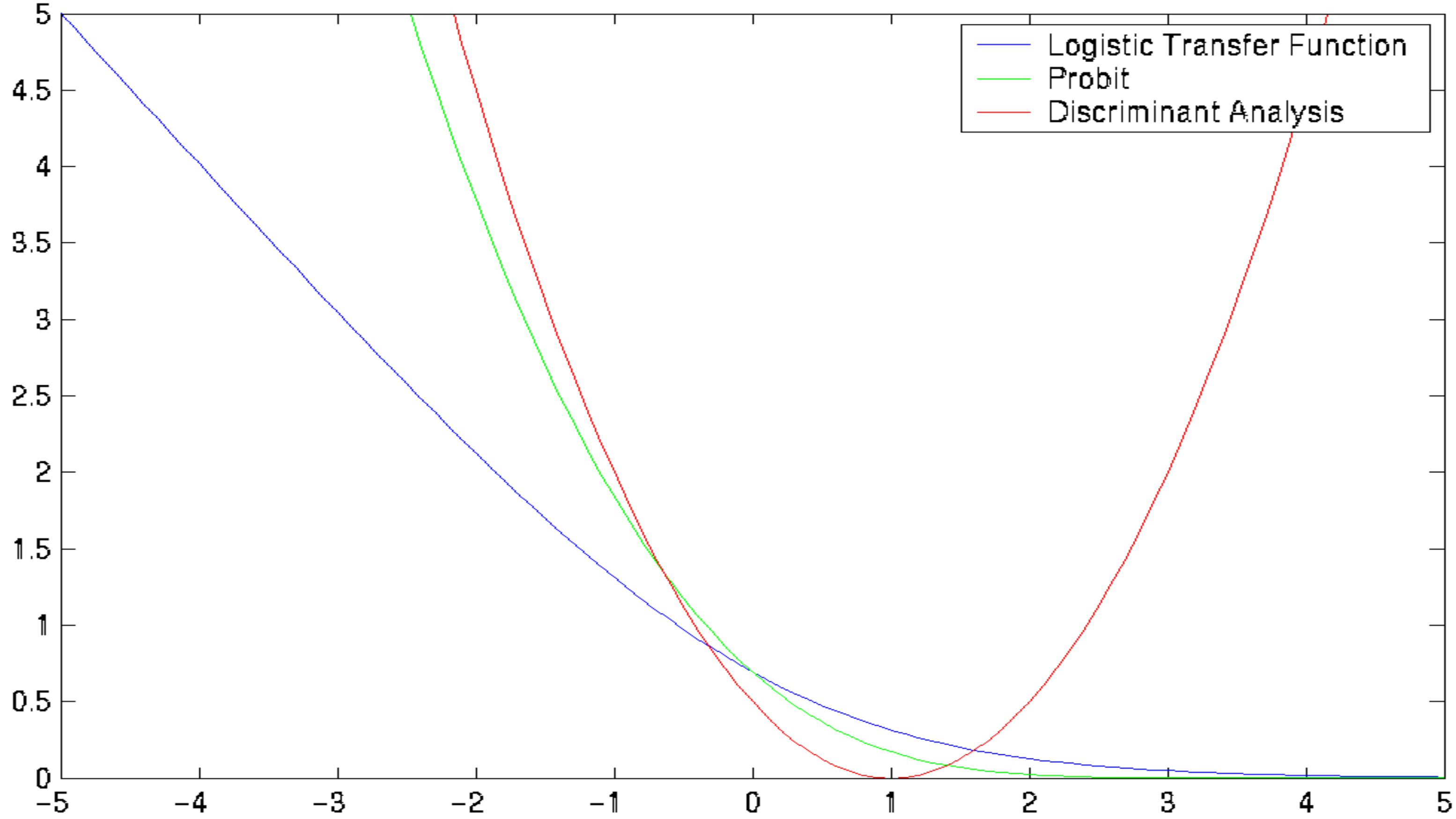
Prediction

For a new instance we obtain $f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x)$ and subsequently predict $y = 1$ if $f(x) > 0$ and $y = -1$ otherwise.

Confidence Ratings

For each observation we get $p(y = 1|x, y) = \frac{1}{1+\exp(f(x))}$.

Link Functions



Soft Margin Loss

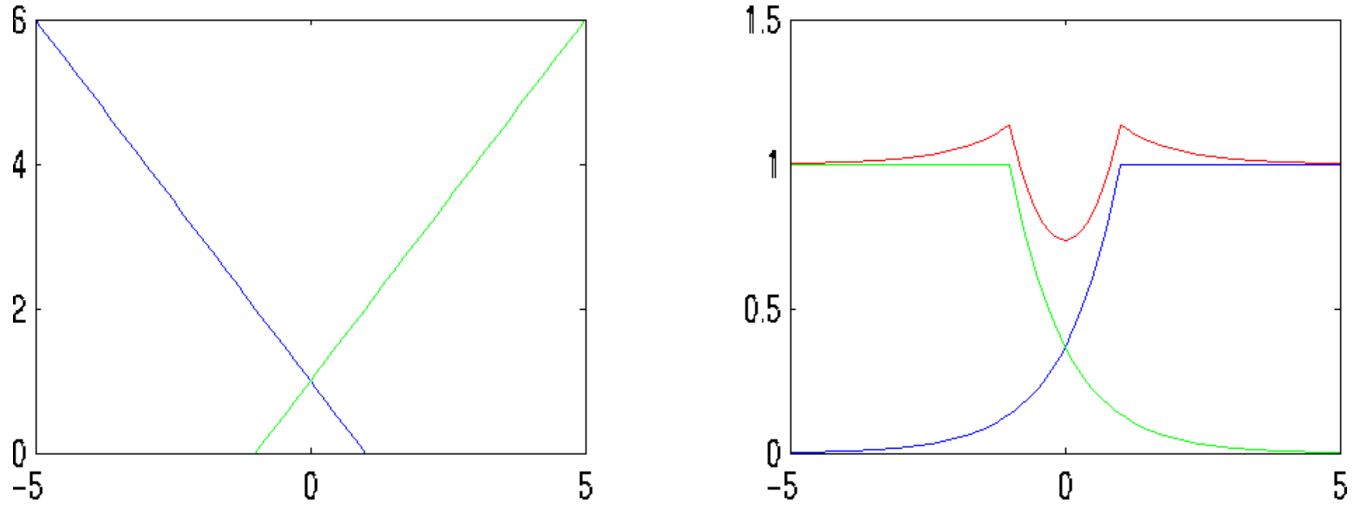
Support Vector Loss Function: In SVM one uses as a loss function

$$c(x, y, f(x)) = \max(0, 1 - yf(x))$$

Using the correspondence between loss functions and log-likelihood, we would get

$$p(y|x, y, f(x)) = \exp(-\max(0, 1 - yf(x))) = \min(1, \exp(yf(x) - 1))$$

Problem: Probabilities don't sum up to 1.



How to fix it

Idea 1

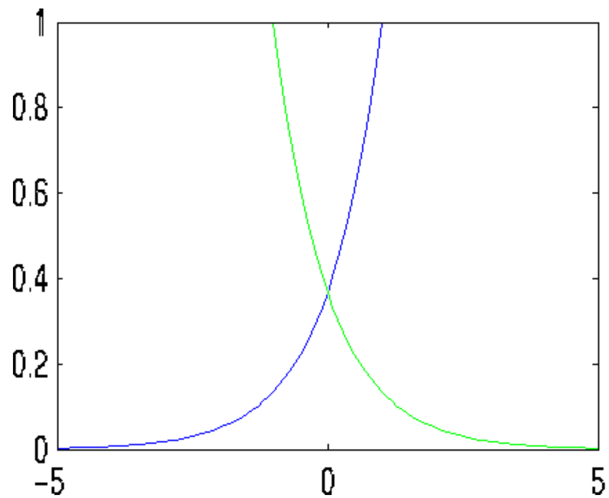
Introduce a "Don't Know" class. This makes sense inside the margin, since we may not be sure which label we have …

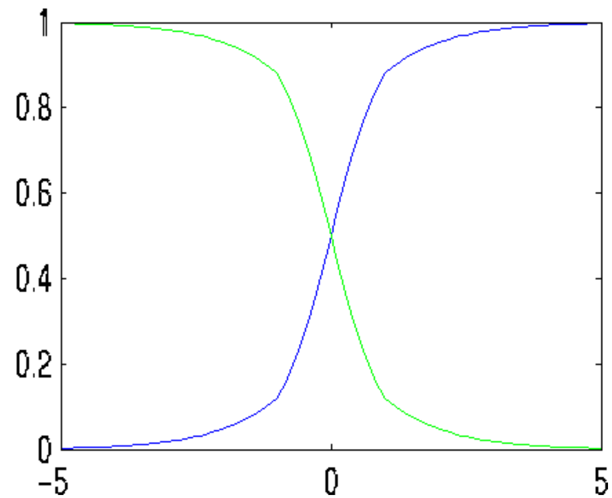
Problem

The "Don't Know" class increases again for large $|f(x)|$. This does not make sense.

Idea 2

Ignore all don't know elements and re-normalize to 1.





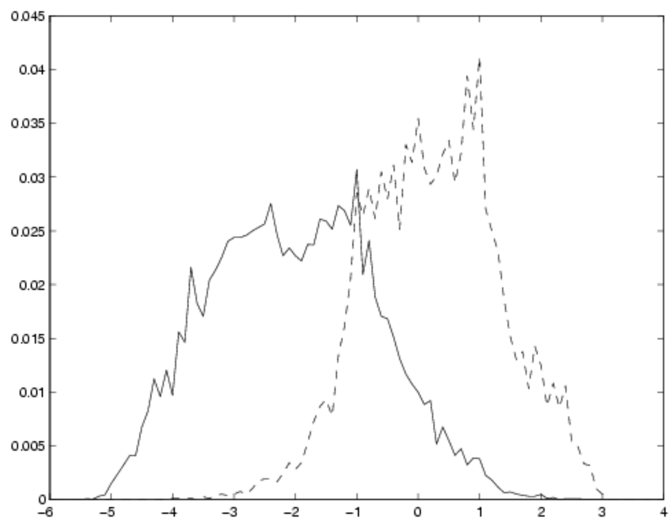
Platt's Trick

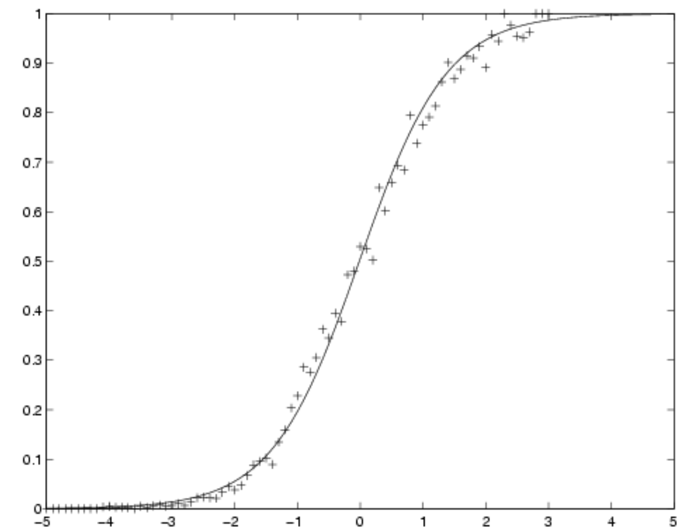
Problem

After obtaining an estimator with a Support Vector Machine we would like to have probabilities (of course, we could have trained a GP estimator straight away) ...

Solution Fit a logistic model to the function values $f(x)$, i.e., we

$$\underset{a,b}{\text{maximize}}\, p(Y|f,X) = \prod_{i=1}^{m} \frac{1}{1+\exp(-ay_i f(x_i)+b)}$$





Why all is well (Proof by Graph)