# Bayesian classification

1. Discriminative classification

2. Bayesian logistic regression & Laplace approximation

3. Generative classification

# Likelihood model

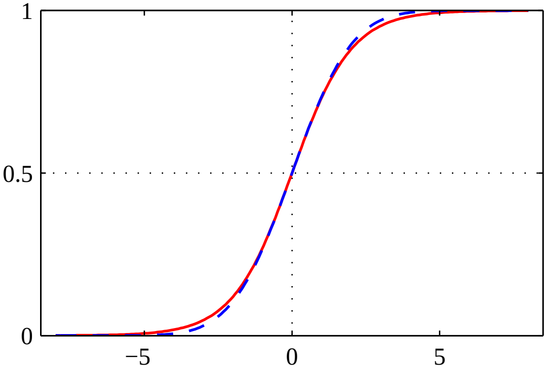
- Two class (binary) classification, discriminative approach: need model for $p(\mathcal{C}_1|\mathbf{x},\mathbf{w}) = 1 - p(\mathcal{C}_2|\mathbf{x},\mathbf{w})$
- Keep this 'almost' linear in parameter vector $\mathbf{w}$:

$$p(\mathcal{C}_1|\mathbf{x},\mathbf{w}) = \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})), \quad \sigma(a) = 1/(1+e^{-a})$$

- $\sigma(a) =$ logistic sigmoid, 'squashing' or 'activation' function
- Inverse: logit $a = \ln(\sigma/(1-\sigma))$, 'link' function
- Model known as 'logistic regression' (but it's classification!)
- Other choices for $\sigma(a)$ are possible, e.g. inverse probit

$$\sigma(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)d\theta$$

# Activation functions




Red: logistic sigmoid; blue: inverse probit

# Likelihood model

- Represent class $\mathcal{C}_1$ as $t = 1$, $\mathcal{C}_2$ as $t = 0$, then

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}, \quad y_n = p(\mathcal{C}_1|\mathbf{x}_n, \mathbf{w}) = \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))$$

- Maximum likelihood minimizes cross-entropy error function

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_n [t_n \ln y_n + (1 - t_n)\ln(1 - y_n)]$$

- Gradient:

$$\nabla E(\mathbf{w}) = \sum_n (y_n - t_n)\boldsymbol{\phi}(\mathbf{x}_n) = \boldsymbol{\Phi}^{\mathrm{T}}(\mathbf{y} - \mathbf{t})$$

# Likelihood model (2)

- Hessian of $E$:

$$\nabla\nabla E(\mathbf{w}) = \sum_n y_n(1-y_n)\boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{R}\boldsymbol{\Phi}$$

with $\mathbf{R}$ = diagonal matrix, $R_{nn} = y_n(1-y_n)$

- Positive definite $\Rightarrow$ $E$ is convex, only has a single minimum
- So $p(\mathbf{t}|\mathbf{w})$ is log-concave, only has a single maximum
- Can be found efficiently numerically (iterative reweighted least squares)

# Generalizations

- Allowing labelling noise:

$$\begin{aligned} p(\mathcal{C}_1|\mathbf{x},\mathbf{w}) &= (1-\epsilon)\sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})) + \epsilon[1-\sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}))] \\ &= \epsilon + (1-2\epsilon)\sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})) \end{aligned}$$

- Classification into $K > 2$ classes: use 'softmax'

$$p(\mathcal{C}_k|\mathbf{x},\mathbf{w}_1\ldots\mathbf{w}_K) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}, \quad a_k = \mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$$

- Likelihood for 1-of-$K$ coding $\mathbf{t}_n$ is then

$$p(\mathbf{t}_1\ldots\mathbf{t}_N|\mathbf{w}_1\ldots\mathbf{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

with $y_{nk} = \exp(a_{nk})/\sum_j \exp(a_{nj})$ and $a_{nk} = \mathbf{w}_k^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)$

# Bayesian classification

# Prior and posterior

- Need to put a prior on $\mathbf{w}$; could choose as for linear regression $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$
- Gives for posterior $p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{w})p(\mathbf{w})$

$$\begin{aligned} \ln p(\mathbf{w}|\mathbf{t}) &= -E(\mathbf{w}) + \text{const.} \\ E(\mathbf{w}) &= \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_n [t_n \ln y_n + (1 - t_n)\ln(1 - y_n)] \end{aligned}$$

- Need to normalize and then integrate to get predictions

$$p(\mathcal{C}_1|\mathbf{x}, \mathbf{t}) = \int p(\mathcal{C}_1|\mathbf{x}, \mathbf{w})p(\mathbf{w}|\mathbf{t})d\mathbf{w}$$

- Not a Gaussian integral – but $p(\mathbf{w}|\mathbf{t})$ has a single maximum
- So approximate by a Gaussian around this maximum: Laplace approximation

# Laplace approximation

## In one dimension

- Consider a generic $p(w) = \exp[-E(w)]/Z$, $Z =$ normalization constant ('partition function')
- If $p(w)$ has a single maximum at $w_0$, can expand around there:

$$E(w) \approx E(w_0) + \frac{1}{2}E''(w_0)(w - w_0)^2$$

- Gives Gaussian approximation for $p(w)$:

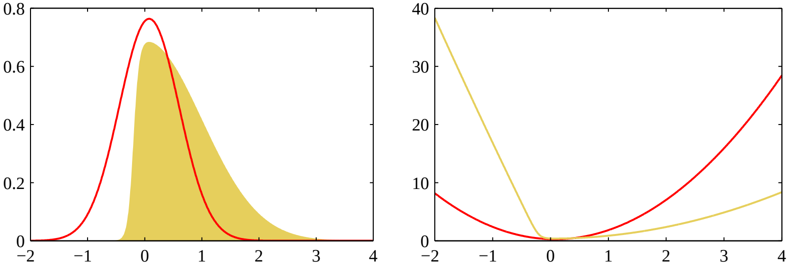
$$p(w) \approx q(w) = \frac{e^{-E(w_0)}}{Z} e^{-\frac{E''(w_0)}{2}(w-w_0)^2} = \mathcal{N}(w|w_0, 1/E''(w_0))$$

- Approximation for $Z$:

$$Z = e^{-E(w_0)}(2\pi)^{1/2}[E''(w_0)]^{-1/2}$$

# Laplace approximation

## Illustration



# Laplace approximation

## In $M$ dimensions

- Consider again $p(\mathbf{w}) = \exp[-E(\mathbf{w})]/Z$
- If $p(\mathbf{w})$ has a single maximum at $\mathbf{w}_0$, can expand around there:

$$E(\mathbf{w}) \approx E(\mathbf{w}_0) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{w} - \mathbf{w}_0)$$

  where $\mathbf{A} = \nabla\nabla\, E(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_0}$ = Hessian at minimum of $E$
- Gives Gaussian approximation for $p(\mathbf{w})$:

$$p(\mathbf{w}) \approx q(\mathbf{w}) = \frac{e^{-E(\mathbf{w}_0)}}{Z} e^{-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{w}-\mathbf{w}_0)} = \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \mathbf{A}^{-1})$$

- Approximation for $Z$:

$$Z = e^{-E(\mathbf{w}_0)}(2\pi)^{M/2}|\mathbf{A}|^{-1/2}$$

# Back to Bayesian logistic regression

- Posterior $p(\mathbf{w}|\mathbf{t}) = \exp[-E(\mathbf{w})]/Z$ with

$$E(\mathbf{w}) = \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_n [t_n \ln y_n + (1-t_n)\ln(1-y_n)]$$

- $E(\mathbf{w})$ convex, single minimum, Hessian $\alpha\mathbf{I} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{\Phi}$
- Find minimum $\mathbf{w}_{\mathrm{MAP}}$, call Hessian there $\mathbf{S}_N^{-1}$
- Then Laplace approximation for posterior is

$$p(\mathbf{w}|\mathbf{t}) \approx q(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{MAP}}, \mathbf{S}_N)$$

# Predictive distribution

- Use approximate posterior:

$$p(\mathcal{C}_1|\mathbf{x},\mathbf{t}) \approx \int p(\mathcal{C}_1|\mathbf{x},\mathbf{w})q(\mathbf{w})d\mathbf{w} = \int \sigma(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}))q(\mathbf{w})d\mathbf{w}$$

- Call $a = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$, then $a$ has Gaussian distribution, with

$$\begin{aligned}
\mathbb{E}[a] &= \int \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\,q(\mathbf{w})d\mathbf{w} = \mathbf{w}_{\mathrm{MAP}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) \\
\mathbb{E}[a^2] &= \int \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w}\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\,q(\mathbf{w})d\mathbf{w} \\
&= \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}(\mathbf{w}_{\mathrm{MAP}}\mathbf{w}_{\mathrm{MAP}}^{\mathrm{T}} + \mathbf{S}_N)\boldsymbol{\phi}(\mathbf{x})
\end{aligned}$$

- So

$$p(\mathcal{C}_1|\mathbf{x},\mathbf{t}) \approx \int \sigma(a)\mathcal{N}(a|\mathbf{w}_{\mathrm{MAP}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))\,da$$

- Can be done numerically, or analytically for inverse probit

# Bayesian classification

1. Discriminative classification
2. Bayesian logistic regression & Laplace approximation
3. Generative classification

# Generative classification

- We model joint distribution $p(\mathbf{x}, \mathcal{C}_k)$, rather than conditional distribution $p(\mathcal{C}_k|\mathbf{x})$ of class labels
- Normally separate $p(\mathbf{x}, \mathcal{C}_k) = p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$
- Class probabilities $p(\mathcal{C}_k|\boldsymbol{\pi}) = \pi_k$
- Class conditional densities e.g.

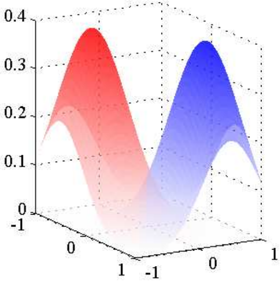
$$p(\mathbf{x}|\mathcal{C}_k, \{\boldsymbol{\mu}_j\}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$$

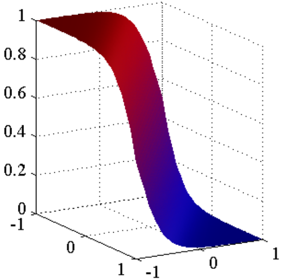
- For two classes this gives

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0)$$

- Linear discriminant as before, logistic sigmoid arises naturally
- If classes have different $\boldsymbol{\Sigma}$, get quadratic discriminant

# Illustration



# Maximum likelihood inference

- Consider two classes, so that $\pi_1 \equiv \pi$, $\pi_2 = 1 - \pi$
- Training data: $N$ inputs $\mathbf{x}_n$, $N$ outputs $t_n \in \{0, 1\}$
- Collect into $\mathbf{X}^{\mathrm{T}} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ and $\mathbf{t} = (t_1, \ldots, t_N)$
- $t_n = 1$ for $\mathcal{C}_1$, $t_n = 0$ for $\mathcal{C}_2$
- Likelihood: $p(\mathbf{x}, t = 1) = p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) = \pi\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$
- Similarly, $p(\mathbf{x}, t = 0) = p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2) = (1 - \pi)\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$
- Overall data likelihood $p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) =$

$$\prod_{n=1}^{N} \left[\pi\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\right]^{t_n} \left[(1 - \pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})\right]^{1-t_n}$$

- Can be maximized in closed form

# Bayesian inference

- Allow $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ now so each $p(\mathbf{x}|\mathcal{C}_k)$ has its own parameters
- Likelihood factorizes: $p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) =$

$$\prod_{n=1}^{N} \pi^{t_n}(1-\pi)^{1-t_n} \prod_{n:t_n=1} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \prod_{n:t_n=0} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$$

- So if prior factorizes into $p(\pi)p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, then posterior $p(\pi, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2|\mathbf{t}, \mathbf{X})$ factorizes in the same way
- Predictive distributions simplify accordingly, e.g. $p(\mathbf{x}, \mathcal{C}_1) =$

$$\int d\pi\, \pi\, p(\pi|\mathbf{t}, \mathbf{X}) \times \int d\boldsymbol{\mu}_1 d\boldsymbol{\Sigma}_1 \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1|\mathbf{t}, \mathbf{X})$$

- Effectively, each class density models $p(\mathbf{x}|\mathcal{C}_k)$ is learnt separately from training data with class label $k$
- Conjugate priors $p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$: Gamma-Wishart