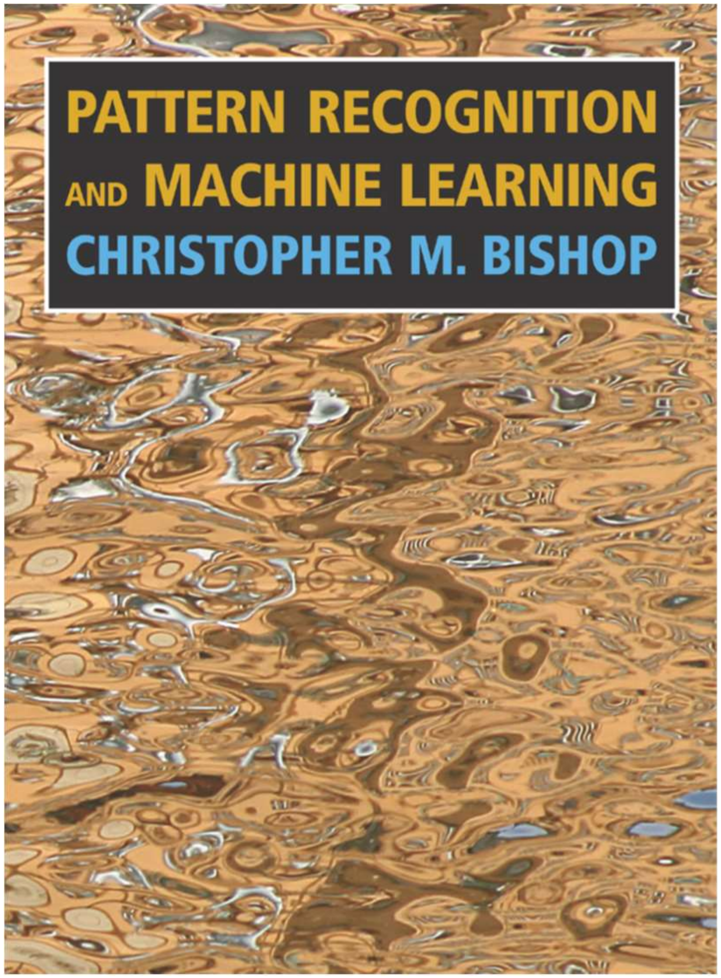
Discriminative Classifiers

Song Liu (song.liu@bristol.ac.uk)

Reference



Today's class *roughly* follows Chapter 4.3

Pattern Recognition and Machine Learning

Christopher Bishop, 2006

Discriminative Classifier

- **Target**: infer $p(y|\boldsymbol{x})$ given dataset D.

- **Step 1.** Making a model assumption $p(y|\boldsymbol{x};\boldsymbol{w})$.
- **Step 2.** Construct the likelihood function $p(D|\boldsymbol{w})$.
- **Step 3.** Estimate the parameters: MLE, MAP, Full Prob…

- **First Question**: What model should we use?

- MVN? NO, that is for continuous variable.
- Our output y is clearly a discrete value.

Modelling $p(y|\boldsymbol{x})$

- Can we express $p(y|\boldsymbol{x})$ using $p(\boldsymbol{x}|y)$?

- Bayes rule says:
- $p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)p(y)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|y)p(y)}{\sum_{y'} p(\boldsymbol{x},y')} = \frac{p(\boldsymbol{x}|y)p(y)}{\sum_{y'} p(\boldsymbol{x}|y')p(y')}$ so

Marginalization!

- Suppose $y \in \{-1,1\}$
- $p(y = 1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y = 1)p(y=1)}{p(\boldsymbol{x}|y' = 1)p(y'=1)+p(\boldsymbol{x}|y' = -1)p(y'=-1)}$

Modelling $p(y|\boldsymbol{x})$

- Suppose $y \in \{-1,1\}$
- $p(y=1|\boldsymbol{x}) = \dfrac{p(\boldsymbol{x}|y=1)p(y=1)}{p(\boldsymbol{x}|y'=1)p(y'=1) + p(\boldsymbol{x}|y'=-1)p(y'=-1)}$
- Nothing has changed, but we are representing $p(y|\boldsymbol{x})$ using $p(\boldsymbol{x}|y)$.
- Assume: $p(\boldsymbol{x}|y)p(y) > 0, \forall \boldsymbol{x}, y$.
- $\dfrac{p(\boldsymbol{x}|y=1)p(y=1)}{p(\boldsymbol{x}|y=1)p(y=1)+p(\boldsymbol{x}|y=-1)p(y=-1)} = \dfrac{1}{1+\boxed{\dfrac{p(\boldsymbol{x}|y=-1)p(y=-1)}{p(\boldsymbol{x}|y=1)p(y=1)}}}$

Modelling $p(y|\boldsymbol{x})$

- We can rewrite $p(y|\boldsymbol{x})$ using the ratio $\frac{p(\boldsymbol{x}|y=-1)p(y=-1)}{p(\boldsymbol{x}|y=1)p(y=1)}$:
- $p(y=1|\boldsymbol{x}) = \frac{1}{1+\frac{p(\boldsymbol{x}|y=-1)p(y=-1)}{p(\boldsymbol{x}|y=1)p(y=1)}}$
- This derivation shows an important difference between generative/discriminative modelling:
- Generative learning models class density $p(\boldsymbol{x}|y)$
- Discriminative learning models density ratio $\frac{p(\boldsymbol{x}|y=-1)}{p(\boldsymbol{x}|y=1)}$!

Modelling Density Ratio

- Clearly, modelling density ratio $\frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=-1)}$ requires a whole lot less assumptions on your class densities.

- Models on $p(\boldsymbol{x}|y) \Rightarrow$ Models $\frac{p(\boldsymbol{x}|y=-1)}{p(\boldsymbol{x}|y=1)}$

- Models on $\frac{p(\boldsymbol{x}|y=-1)}{p(\boldsymbol{x}|y=1)}$ $\not\Rightarrow$ Models $p(\boldsymbol{x}|y)$

Modelling Log-Density Ratio

- $p(y=1|\boldsymbol{x}) = \dfrac{1}{1+\boxed{\dfrac{p(\boldsymbol{x}|y=-1)p(y=-1)}{p(\boldsymbol{x}|y=1)p(y=1)}}}$

  $\Rightarrow p(y=1|\boldsymbol{x},\boldsymbol{w}) := \dfrac{1}{1+\boxed{\exp\left(-f(\boldsymbol{x};\boldsymbol{w})\right.}\left.\right)}$

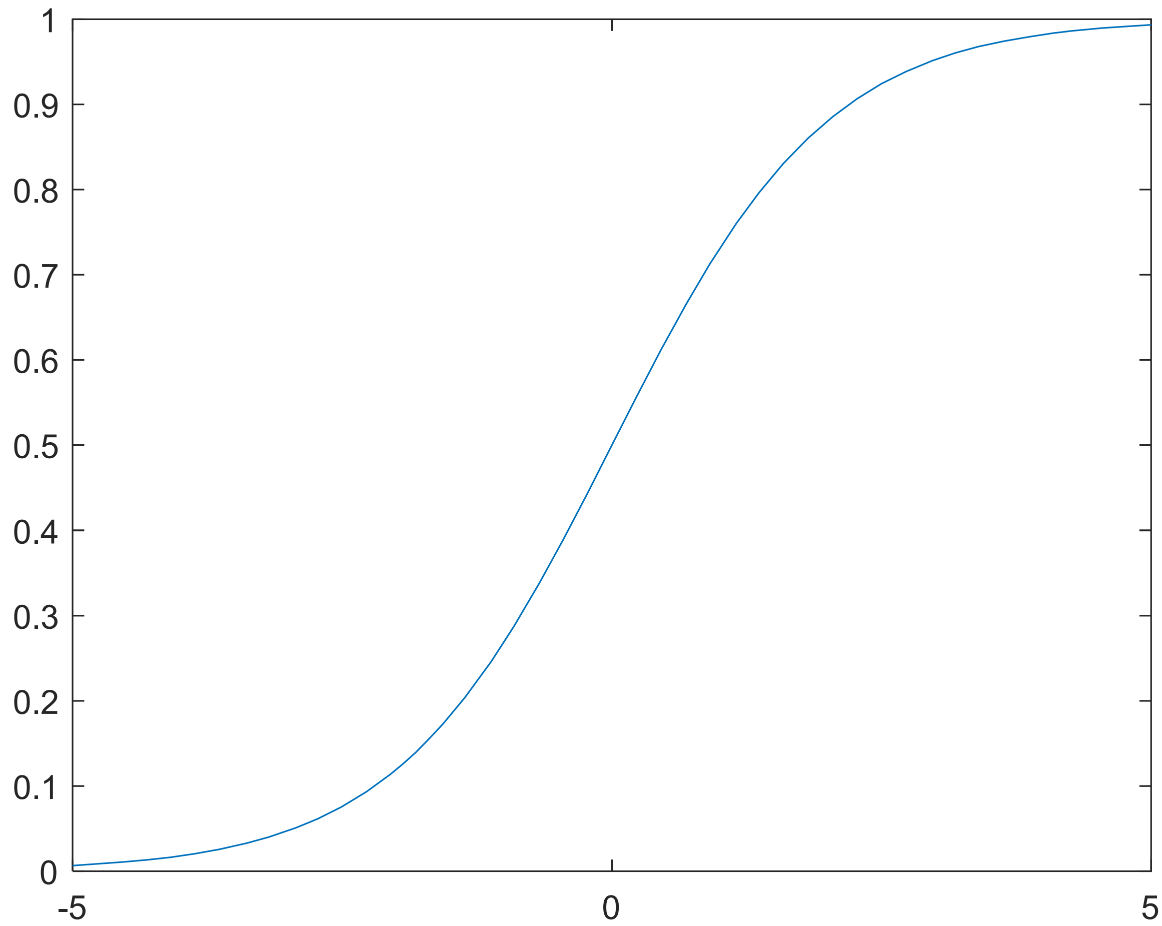
- We model log ratio, $\log \dfrac{p(\boldsymbol{x}|y=1)p(y=1)}{p(\boldsymbol{x}|y=-1)p(y=-1)}$ as $f(\boldsymbol{x};\boldsymbol{w})$

- Like density estimation, it is better to work with log-ratio rath than the ratio itself.

Generalized Linear Model

- As usual, $f(\boldsymbol{x};\boldsymbol{w}) = \langle \boldsymbol{w}', \boldsymbol{x} \rangle + w_0$.
- Let $\sigma(t) \coloneqq \frac{1}{1+\exp(-t)}$, “sigmoid function”
- The model for $p(y|\boldsymbol{x};\boldsymbol{w}) := \sigma\big(f(\boldsymbol{x};\boldsymbol{w})\big)$ is merely a linear function wrapped by a non-linear transform.
- We call $\sigma\big(f(\boldsymbol{x};\boldsymbol{w})\big)$ a “generalized linear model”. This model is widely used in places beyond classification.

Sigmoid Function $\sigma(t) := \dfrac{1}{1+\exp(-t)}$



Modelling Log-Density Ratio

- $p(y=-1|\boldsymbol{x}) = \dfrac{1}{1+\dfrac{p(\boldsymbol{x}|y=+1)p(y=+1)}{p(\boldsymbol{x}|y=-1)p(y=-1)}}$

  $$\Rightarrow p(y=-1|\boldsymbol{x},\boldsymbol{w}) := \frac{1}{1+\exp\big(f(\boldsymbol{x};\boldsymbol{w})\big)}$$

- In $p(y=-1|\boldsymbol{x})$, $\dfrac{p(\boldsymbol{x}|y=+1)p(y=+1)}{p(\boldsymbol{x}|y=-1)p(y=-1)}$ occurs, which is the exact inverse of the ratio appeared in $p(y=1|\boldsymbol{x})$. This ratio is modelled by $\exp\big(f(\boldsymbol{x};\boldsymbol{w})\big)$.

- To simplify our model, let us write
- $p(y|\boldsymbol{x};\boldsymbol{w}) := \sigma\big(f(\boldsymbol{x};\boldsymbol{w})\cdot y\big)$

Estimate $p(y|\boldsymbol{x}; \boldsymbol{w})$ from D

- Assuming the IID-ness on D.
- Likelihood: $p(D|\boldsymbol{w}) = \prod_{i \in D} p(y_i|\boldsymbol{x}_i; \boldsymbol{w})$,
- Just like what we did for regression tasks.

- MLE for $p(y|\boldsymbol{x}; \boldsymbol{w})$:
- $\boldsymbol{w}_{\text{MLE}} = \text{argmax}_{\boldsymbol{w}} \log \prod_{i \in D} p(y_i|\boldsymbol{x}_i; \boldsymbol{w})$

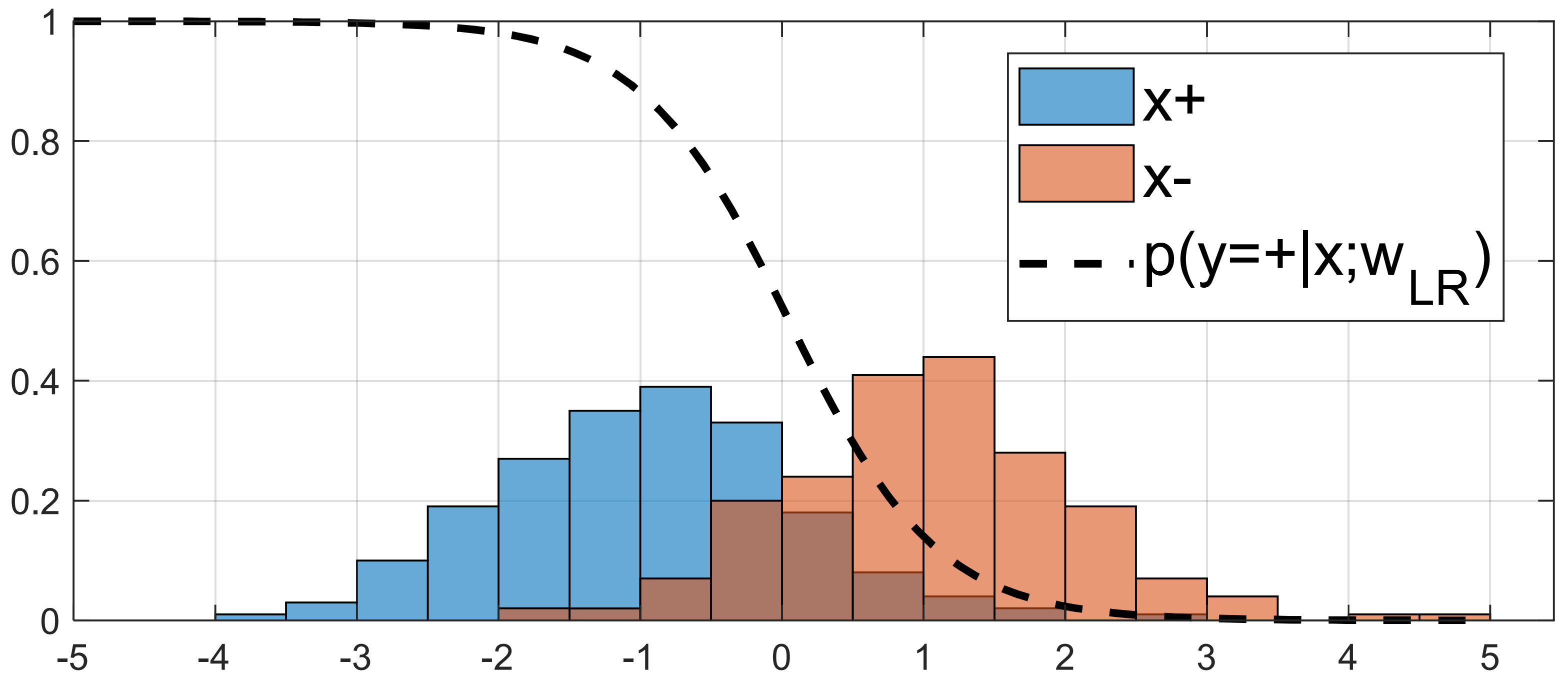
$$= \text{argmax}_{\boldsymbol{w}} \sum_{i \in D} \log p(y_i|\boldsymbol{x}_i; \boldsymbol{w})$$

$$= \text{argmax}_{\boldsymbol{w}} \sum_{i \in D} \log \sigma(f(\boldsymbol{x}_i; \boldsymbol{w}) \cdot y_i)$$

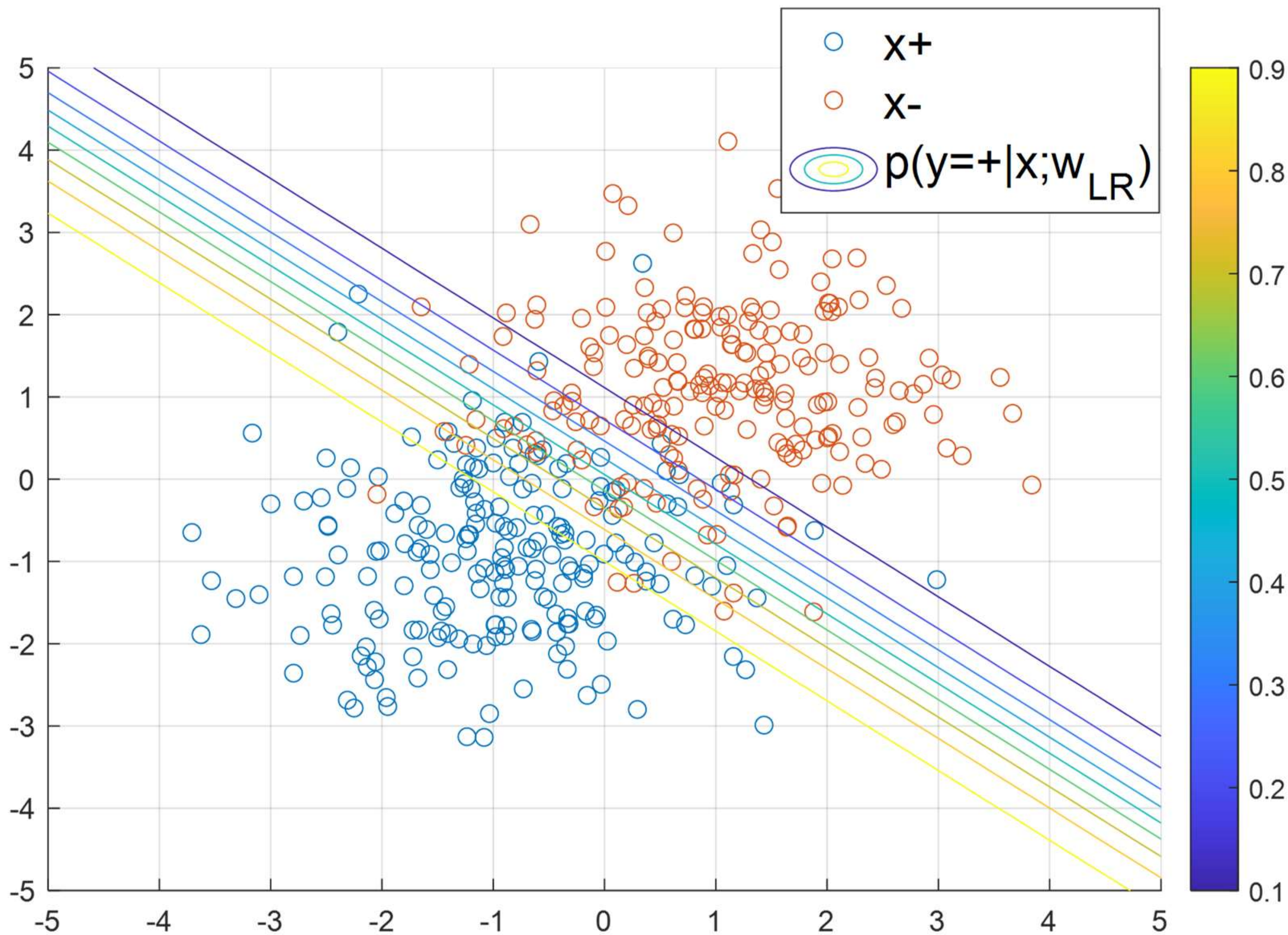
Logistic Regression

- This MLE procedure is also called Logistic Regression.
- **This. Is. Not. A. Regression!**

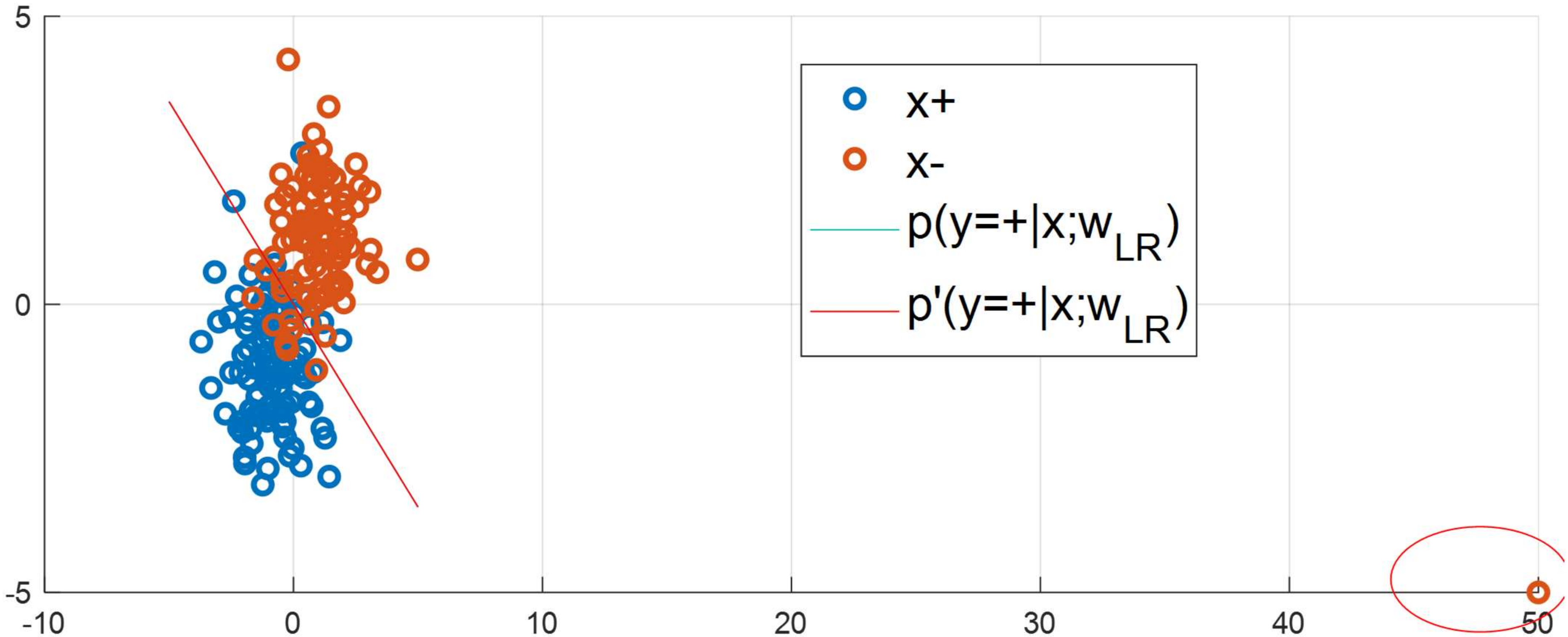
Logistic Regression



Logistic Regression 2D

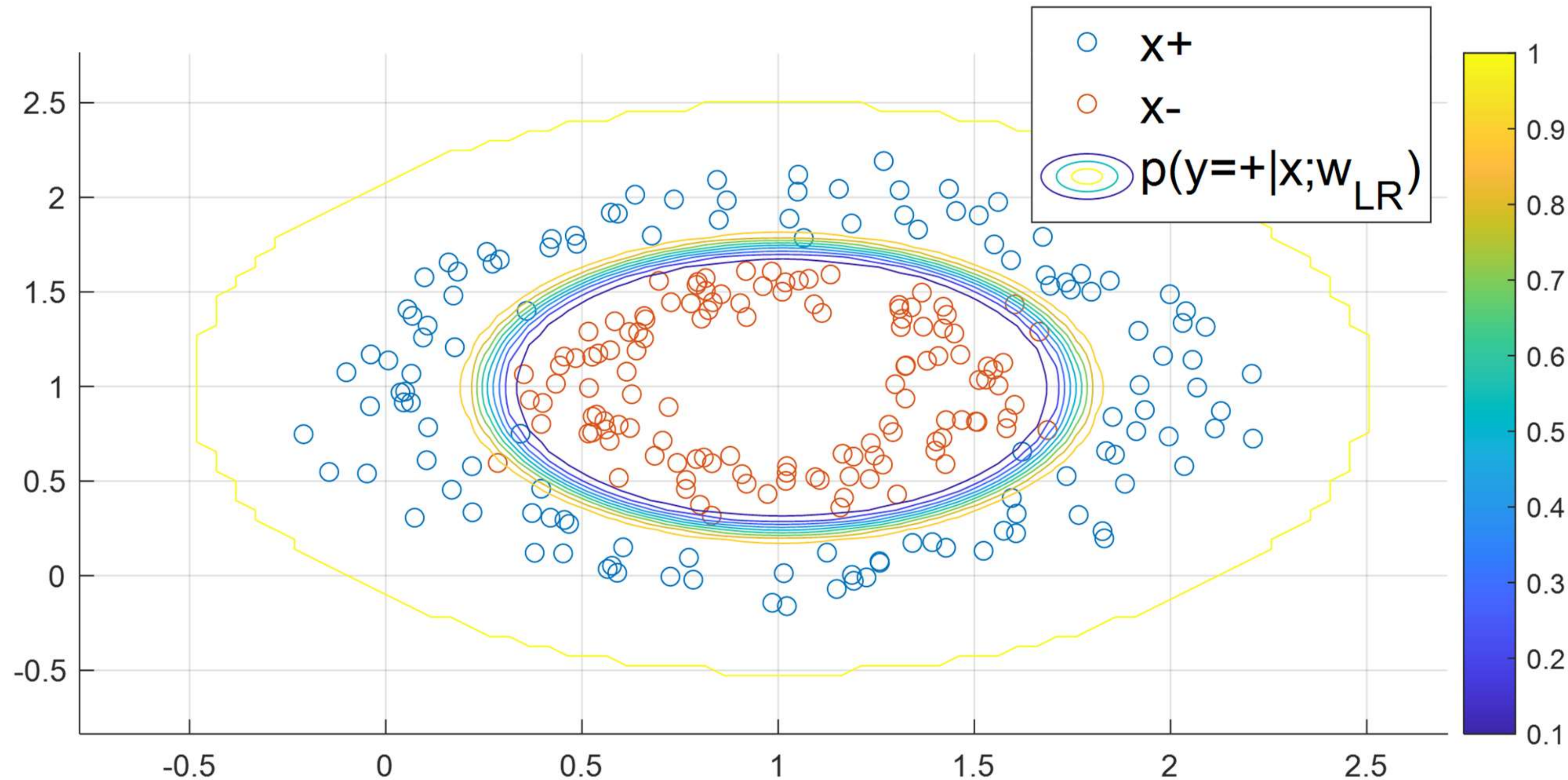


Robustness of Logistic Regression



Unlike LS classifier, LR is not affected by outliers that are far away from the decision boundary. Why?

Logistic Regression with Feature Transform $\boldsymbol{\phi}(\boldsymbol{x})$



- Since $f(\boldsymbol{x};\boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ still takes a linear form, we can replace $\boldsymbol{x}$ with $\boldsymbol{\phi}(\boldsymbol{x})$ to create a non-linear classifier.
- $\boldsymbol{\phi}$ can be Poly. Trignometric, or RBF.

Estimating $p(y|\boldsymbol{x};\boldsymbol{w})$

- We can assume priors on $\boldsymbol{w}$, then
- $\boldsymbol{w}_{\text{MAP}} = \text{argmax}_{\boldsymbol{w}} \sum_{i \in D} \log(\sigma(f(\boldsymbol{x}_i;\boldsymbol{w}) \cdot y_i)\, p(\boldsymbol{w}))$

$$= \text{argmax}_{\boldsymbol{w}} \sum_{i \in D} \log \sigma(f(\boldsymbol{x}_i;\boldsymbol{w}) \cdot y_i) + \log p(\boldsymbol{w})$$

- We can also use the full prob. approach
- $p(y|\boldsymbol{x}) = \int p(y|\boldsymbol{x};\boldsymbol{w}) p(\boldsymbol{w}|D) d\boldsymbol{w}$

$$\propto \int p(y|\boldsymbol{x};\boldsymbol{w}) p(D|\boldsymbol{w}) p(\boldsymbol{w}) d\boldsymbol{w}$$

$$\propto \int \sigma(f(\boldsymbol{x}_i;\boldsymbol{w}) \cdot y_i) \prod_{i \in D} \sigma(f(\boldsymbol{x}_i;\boldsymbol{w}) \cdot y_i)\, p(\boldsymbol{w}) d\boldsymbol{w}$$

- Unlike regression using MVN models, we cannot calculate this integral in closed form. See PRML 4.4, 4.5.

Multi-class Logistic Regression

- It is easy to extend logistic regression to a multi-class classification problem.
- $p(y=1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y=1)p(y=1)}{\boxed{\sum_k p(\boldsymbol{x}|y=k)p(y=k)}}$

Marginalization is no longer with respect to a binary y!

- This expression enables an elegant expression of logistic regression objective using one-hot encoding.

One-hot Logistic Regression

- $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{w}) = \boldsymbol{W}^{\top}\,\widetilde{\boldsymbol{x}}, \mathbf{W} \in R^{d\times K}$, $\widetilde{\boldsymbol{x}} := [\boldsymbol{x}^{\top}, \mathbf{1}]^{\top}$
- Use “**one hot encoding**”: $y_i \in \{1 \dots K\} \Rightarrow \boldsymbol{t}_i \in R^K$

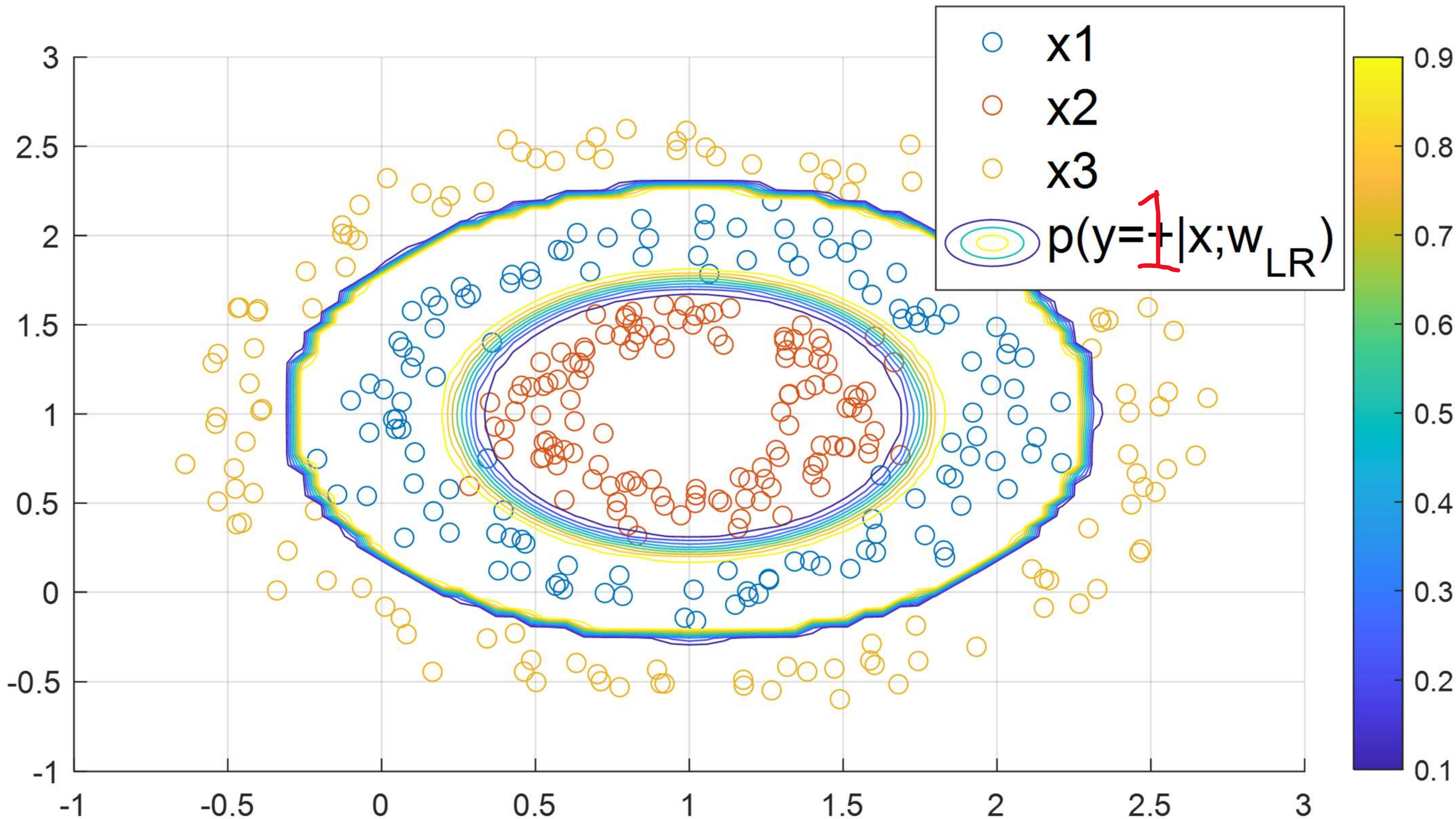
- $\boldsymbol{w}_{\mathrm{MLE}} = \operatorname{argmax}_{\boldsymbol{w}} \sum_{i\in D} \log \sigma(\boldsymbol{f}(\boldsymbol{x}_i;\boldsymbol{w}), \boldsymbol{t}_i)$
- where $\sigma(\boldsymbol{f},\boldsymbol{t}) := \frac{\exp\langle \boldsymbol{f},\boldsymbol{t}\rangle}{\sum_k \exp f^{(k)}}$.
- **Homework: What is the probabilistic interpretation of $\boldsymbol{f}$?**
- If prediction is given by $\operatorname{argmax}_y p(y|\boldsymbol{x};\boldsymbol{W})$, it corresponds to multi-class decision rule we saw in previous lecture. Why?

Multi-class Classification

Previous Lecture

- Rather than relying on sign of f to make predictions, we estimate K functions:
- $\{f_k(\boldsymbol{x}; \boldsymbol{w}_k)\}_{k=1}^{K}$
- Given an $\boldsymbol{x}$, prediction is $\hat{k}$ if $f_{\hat{k}}(\boldsymbol{x}; \boldsymbol{w}_{\hat{k}}) > f_j(\boldsymbol{x}; \boldsymbol{w}_j), \forall j$
- **Problem**: f_k does not have a simple geometry interpretation anymore.
- However, f_k does have probabilistic interpretation.

Multi-class Logistic Regression



Implementation of Logistic Regression

- Unlike LS, LR does not have a closed form solution.

- It means, to find $\boldsymbol{w}_{\mathrm{MLE}}$, we need to solve

$$\mathrm{argmax}_{\boldsymbol{w}} \sum_{i \in D} \log \sigma(f(\boldsymbol{x}_i; \boldsymbol{w}) \cdot y_i)$$

- numerically!!
- The implementation of this algorithm requires some knowledge on numerical optimization, which is not introduced in this class.
- Fortunately, numerical optimization packages are readily available in many programming languages.

Conclusion

- Discriminative classification models **density ratio** while generative classification models **class densities**.
- When log-ratio is modelled by $f(\boldsymbol{x};\boldsymbol{w}) \coloneqq \langle \boldsymbol{w}', \boldsymbol{x} \rangle + w_0$, the model for the class posterior $p(y|x)$ is called generalized linear model.
- The MLE solution for generalized linear model is called logistic regression.
  - whose solution requires numerical optimization.

Homework

- What are the **decision functions** given by a binary logistic regression? (hint: $p(y|\boldsymbol{x};\boldsymbol{w}) - .5$ is one of them)

- Prove: if $p(\boldsymbol{x}|y=1)$ and $p(\boldsymbol{x}|y=-1)$ are MVN with shared covariance matrix $\boldsymbol{\Sigma}$ but different means $\boldsymbol{\mu}_+, \boldsymbol{\mu}_-$.
- 1. $\exists \boldsymbol{w}^*$ such that $p(y|\boldsymbol{x}) = \sigma\big((\langle \boldsymbol{x}; \boldsymbol{w}'^* \rangle + w_0'^*)y\big)$
- 2. find $\boldsymbol{w}^*$

- Show the probabilistic interpretation of multiclass logistic regression

Jensen Shannon Divergence (Challenging)

- Similar to KL divergence, Jensen Shannon divergence is a discrepancy measure between two probability density functions p and q.
- $\mathrm{JS}[p, q] := \frac{1}{2}\mathrm{E}_p\left[\log \frac{p(x)}{.5p(x)+.5q(x)}\right] + \frac{1}{2}\mathrm{E}_q\left[\log \frac{q(x)}{.5p(x)+.5q(x)}\right]$.
- How is the LR objective related to $\mathrm{JS}[p, q]$ when $p(y = 1) = p(y = -1)$?
- Hint: What is the maximiser of the following problem?
- $argmax_t \mathrm{E}_p[\log t(x)] + \mathrm{E}_q[\log(1 - t(x))]$, where t is a function $t: R^d \to R, t \in (0,1)$.