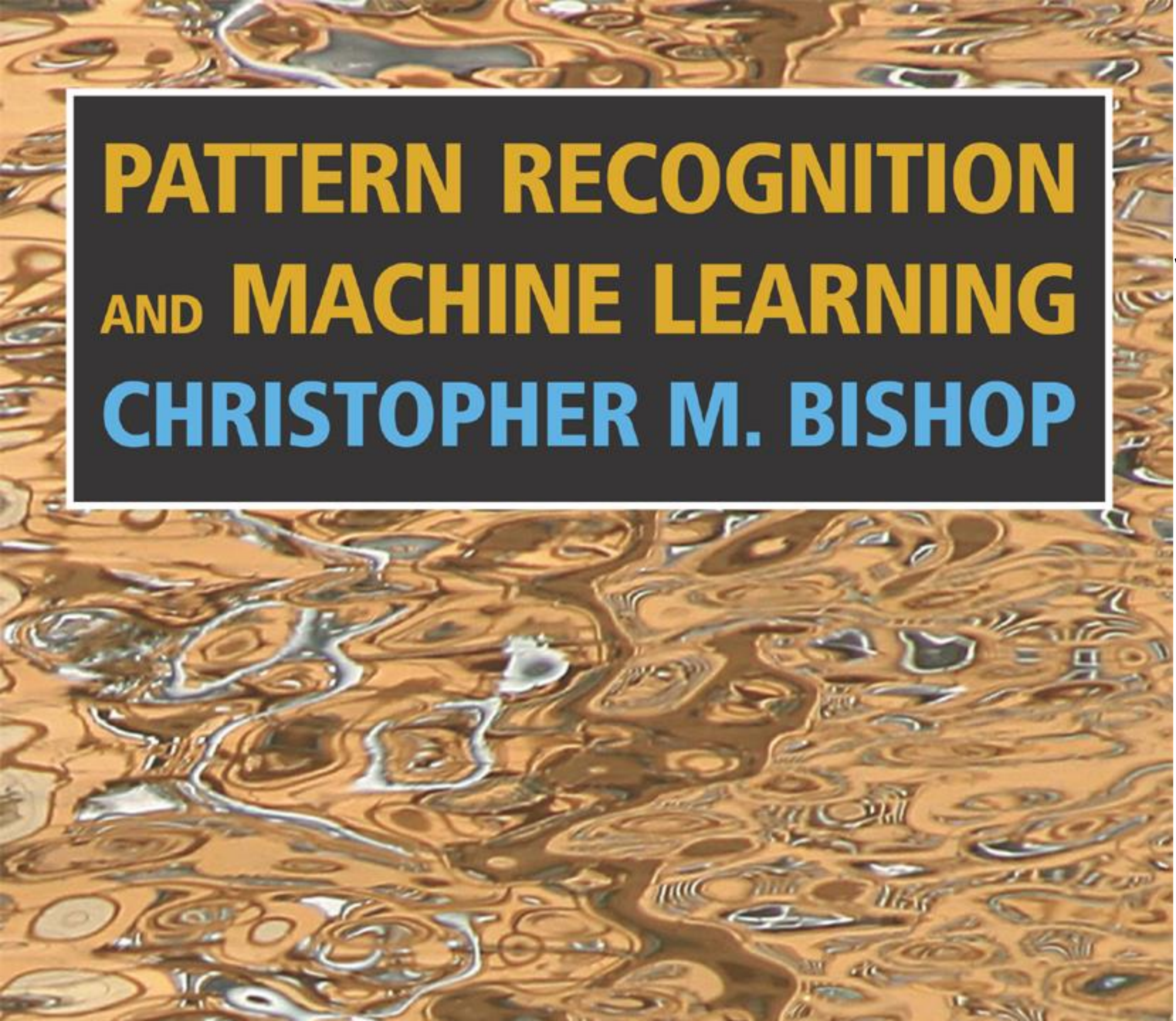
Discriminative Classifiers

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


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP

ws

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

- We check what $p(y|\boldsymbol{x})$ looks like.

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

- Nothing has changed, but we are representing $p(y|\boldsymbol{x})$ using $p(\boldsymbol{x}|y)$.
- Assume: $p(\boldsymbol{x}|y)p(y) > 0, \forall \boldsymbol{x}, y$.

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

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

- After some mild assumptions,
- 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.

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

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

Modelling Log-Density Ratio

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

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

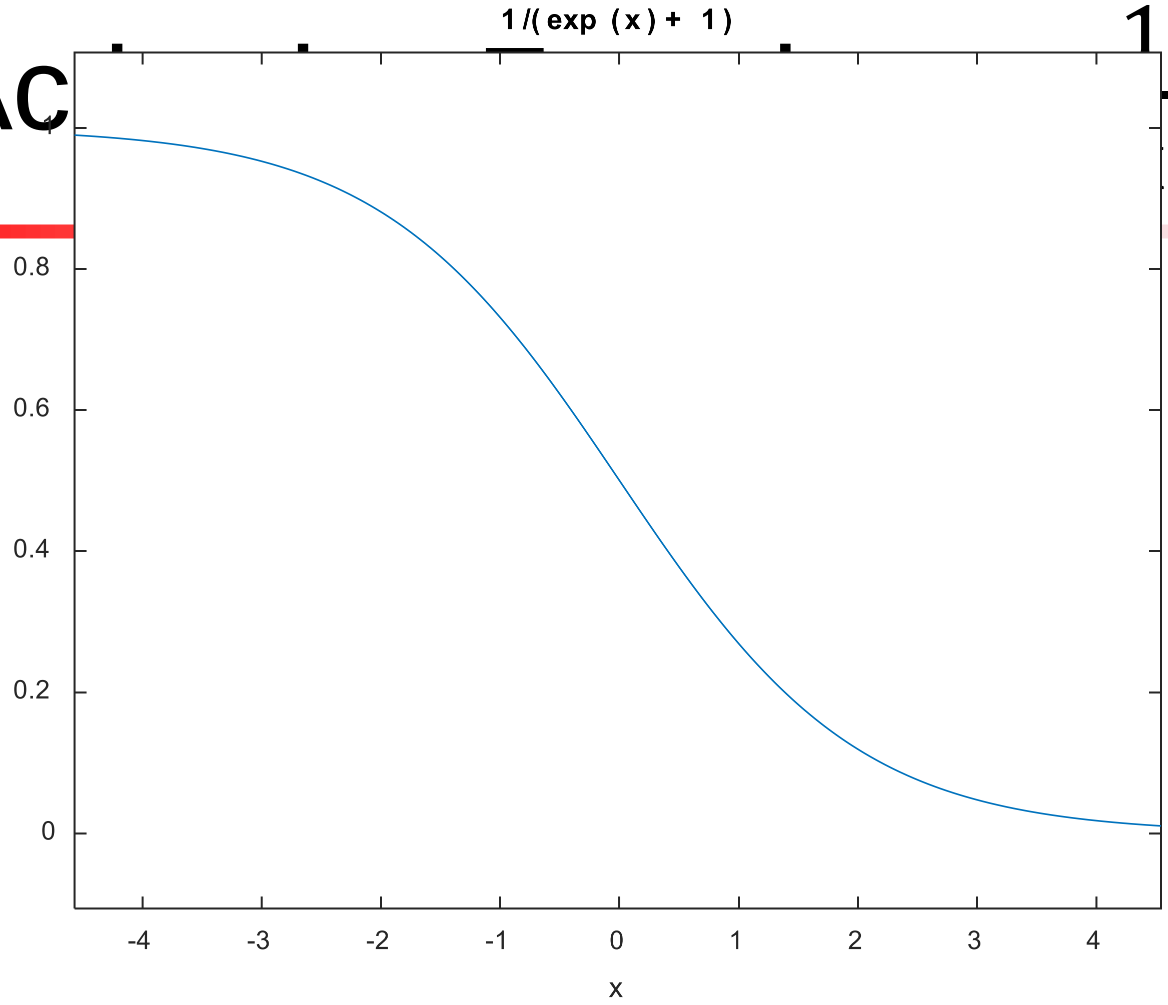
- We model log ratio, $\log \dfrac{p(\boldsymbol{x}|y = -1)p(y=-1)}{p(\boldsymbol{x}|y = 1)p(y=1)}$ as $f(\boldsymbol{x}; \boldsymbol{w})$

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

Generalized Linear Model

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

Ac ... $\frac{1}{\ldots \exp(t)}$



Modelling Log-Density Ratio

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

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

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

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

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

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

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

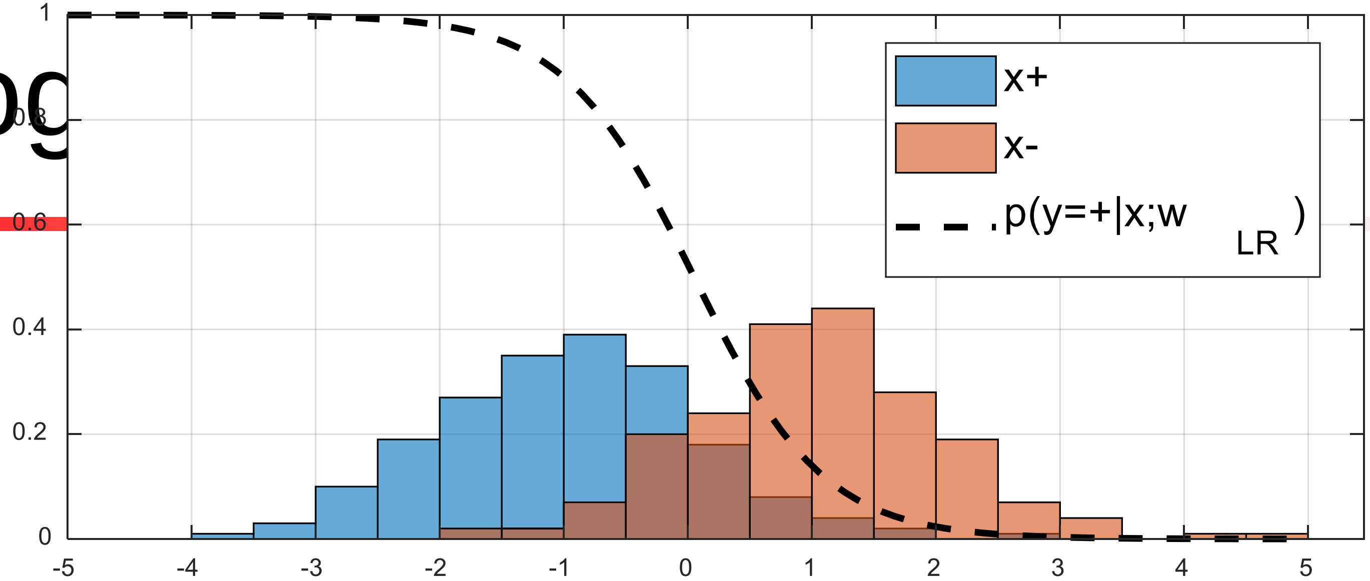
$$= \mathrm{argmax}_{\boldsymbol{w}} \sum_{i\in D} \log p(y_i|\boldsymbol{x}_i;\boldsymbol{w}) \; + C$$

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

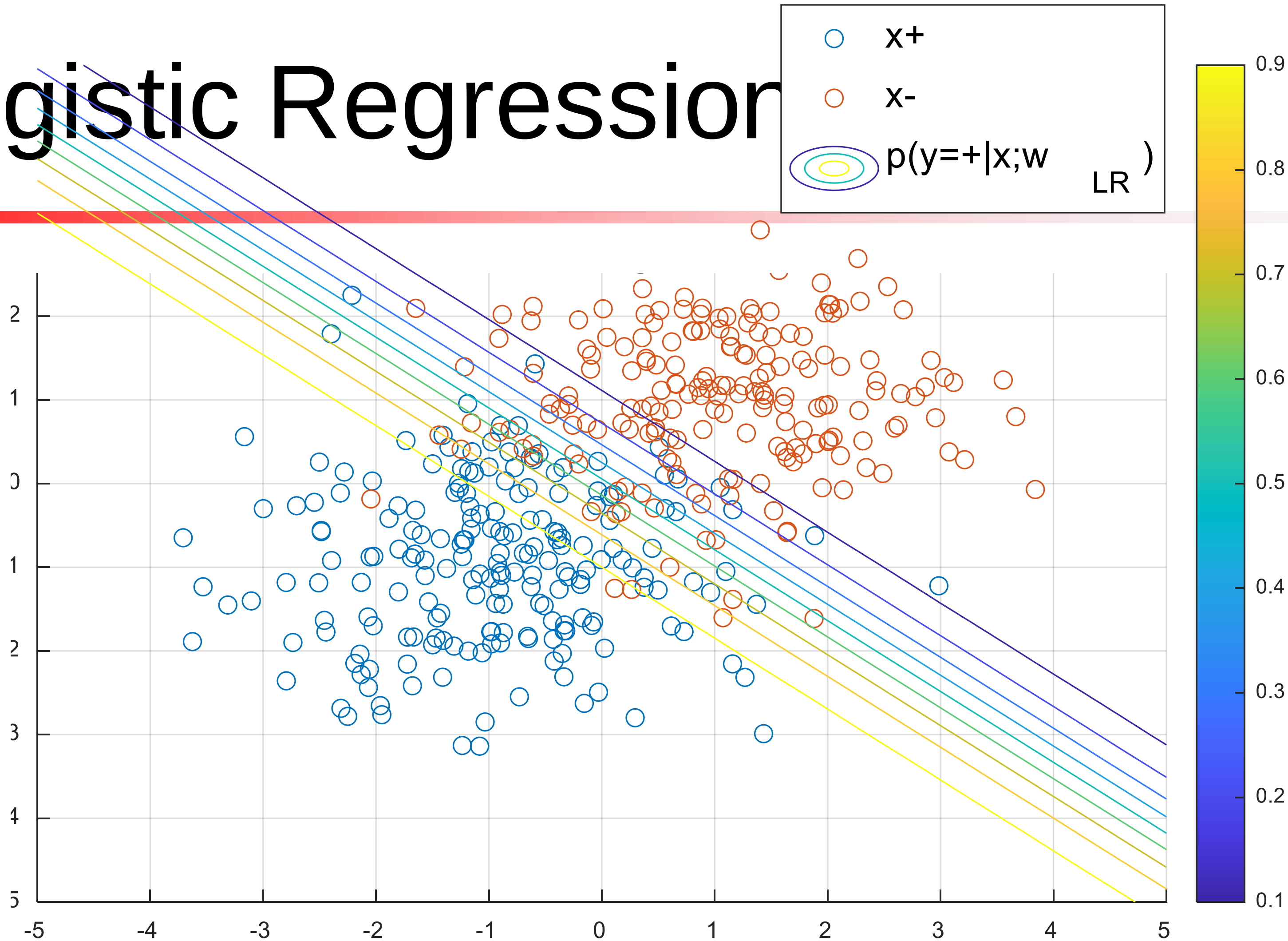
Logistic Regression (rant)

- The MLE solution of $\boldsymbol{w}$ is also called Logistic Regression.
- **This is the worst. name. ever.**
- The first word does not make much sense and the second word misleads you to think this is a regression algorithm while in fact it is a classification algorithm!
  - The "logistic" comes from the fact that σ is also called logistic function in mathematics.
- But everybody still uses this name, so we call it that.

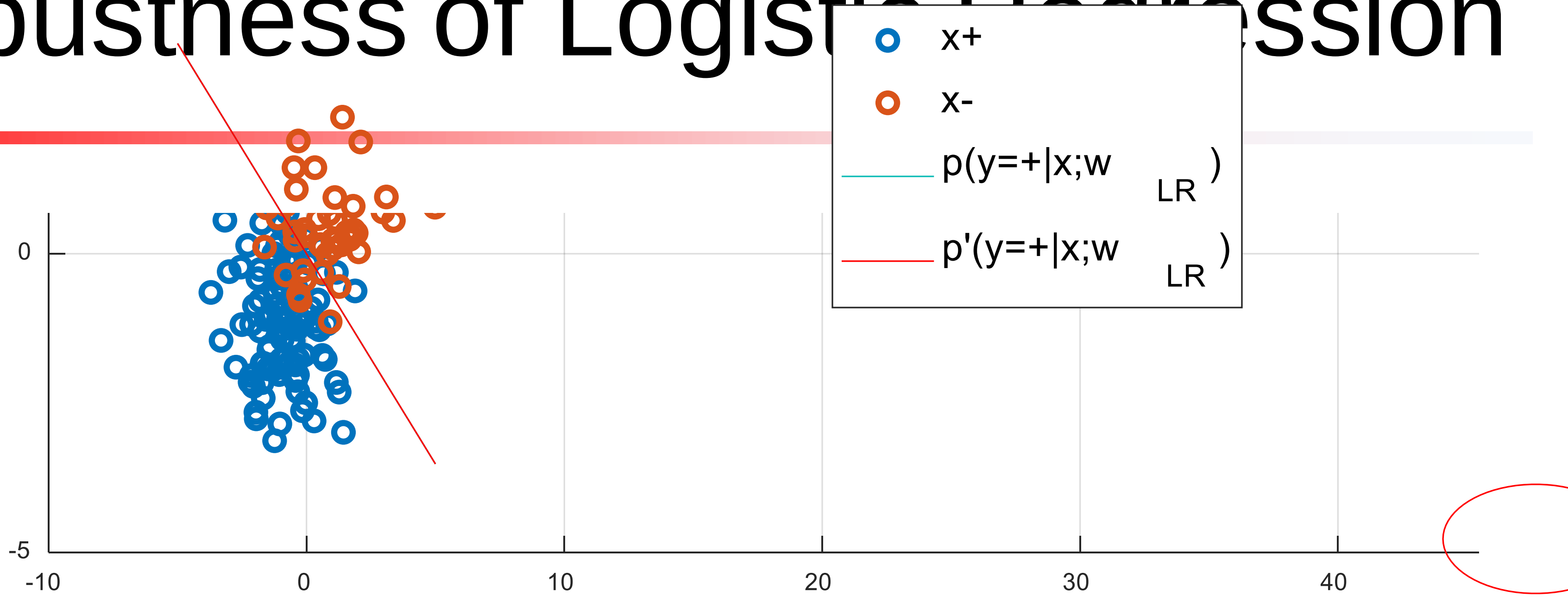
Log



Logistic Regression

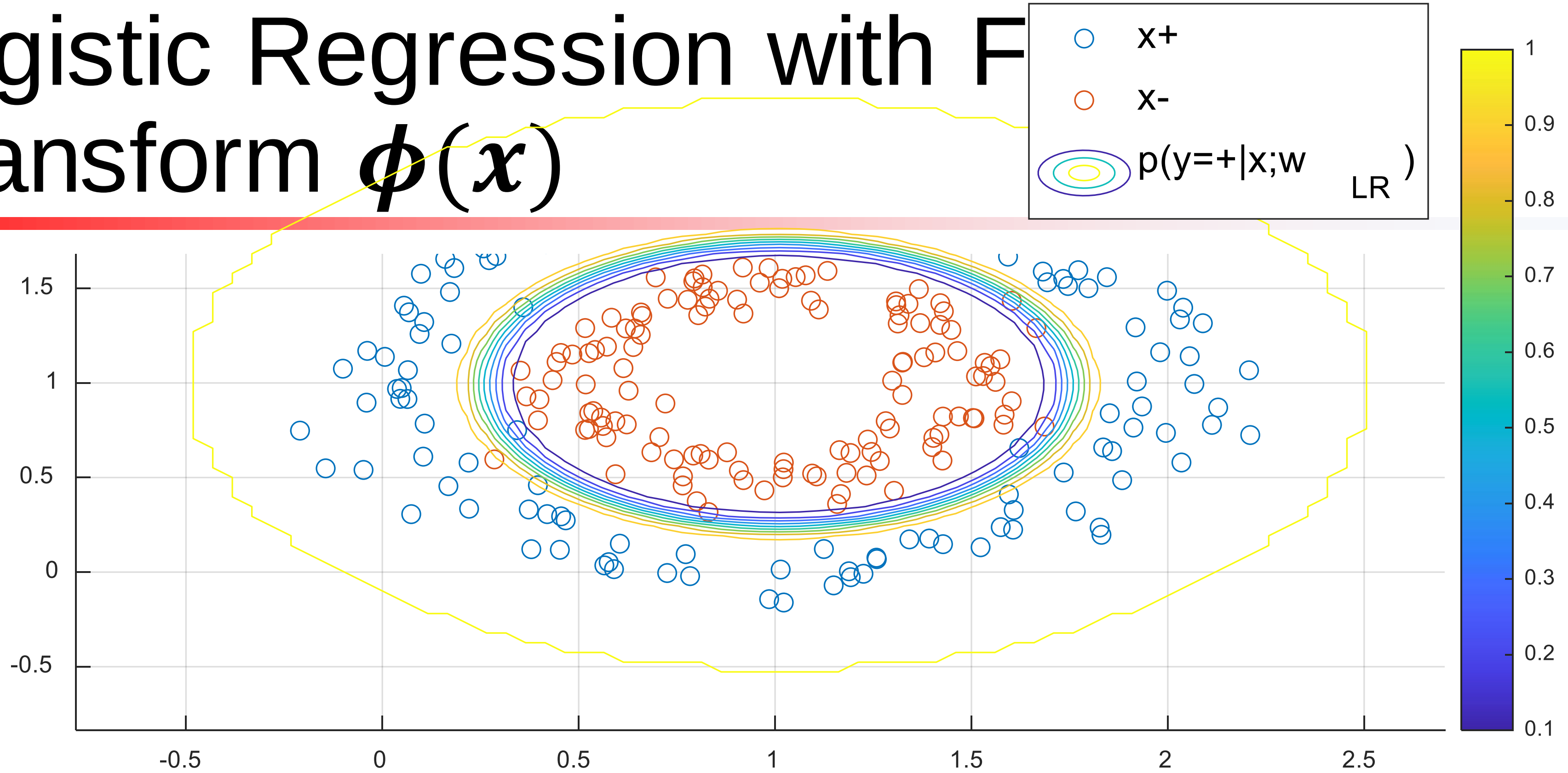


Robustness of Logistic Regression



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

Logistic Regression with F
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.

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

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

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

Marginalization is no longer with respect to a binary K!

- Model the log-ratio, $\log \dfrac{p(\boldsymbol{x}|y=k')p(y=k')}{p(\boldsymbol{x}|y=k)p(y=k)}$ as:
- $f(\boldsymbol{x};\boldsymbol{w}_{k'},\boldsymbol{w}_k) = f'(\boldsymbol{x};\boldsymbol{w}_{k'}) - f'(\boldsymbol{x};\boldsymbol{w}_k)$
- This model allows an elegant expression of logistic regression using one-hot encoding.

One-hot Logistic Regression

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

- $\boldsymbol{w}_{\text{MLE}} = \text{argmax}_{\boldsymbol{w}} \sum_{i\in D} \log \sigma(\boldsymbol{f}(\boldsymbol{x}_i;\boldsymbol{w}), \boldsymbol{t}_i)$
- where $\sigma(\boldsymbol{f},\boldsymbol{t}) := \frac{\exp\langle \boldsymbol{f},\boldsymbol{t}\rangle}{\sum_k \exp f^{(k)}}$.
- This expression uses the log-ratio model we defined in the previous slide. Verify this!

- Here $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{w})$ gives us K prediction functions. If prediction is given by $\text{argmax}_y p(y|\boldsymbol{x};\boldsymbol{W})$, it corresponds to multi-class decision rule we saw in previous lecture. Why?

Multi-class Classification

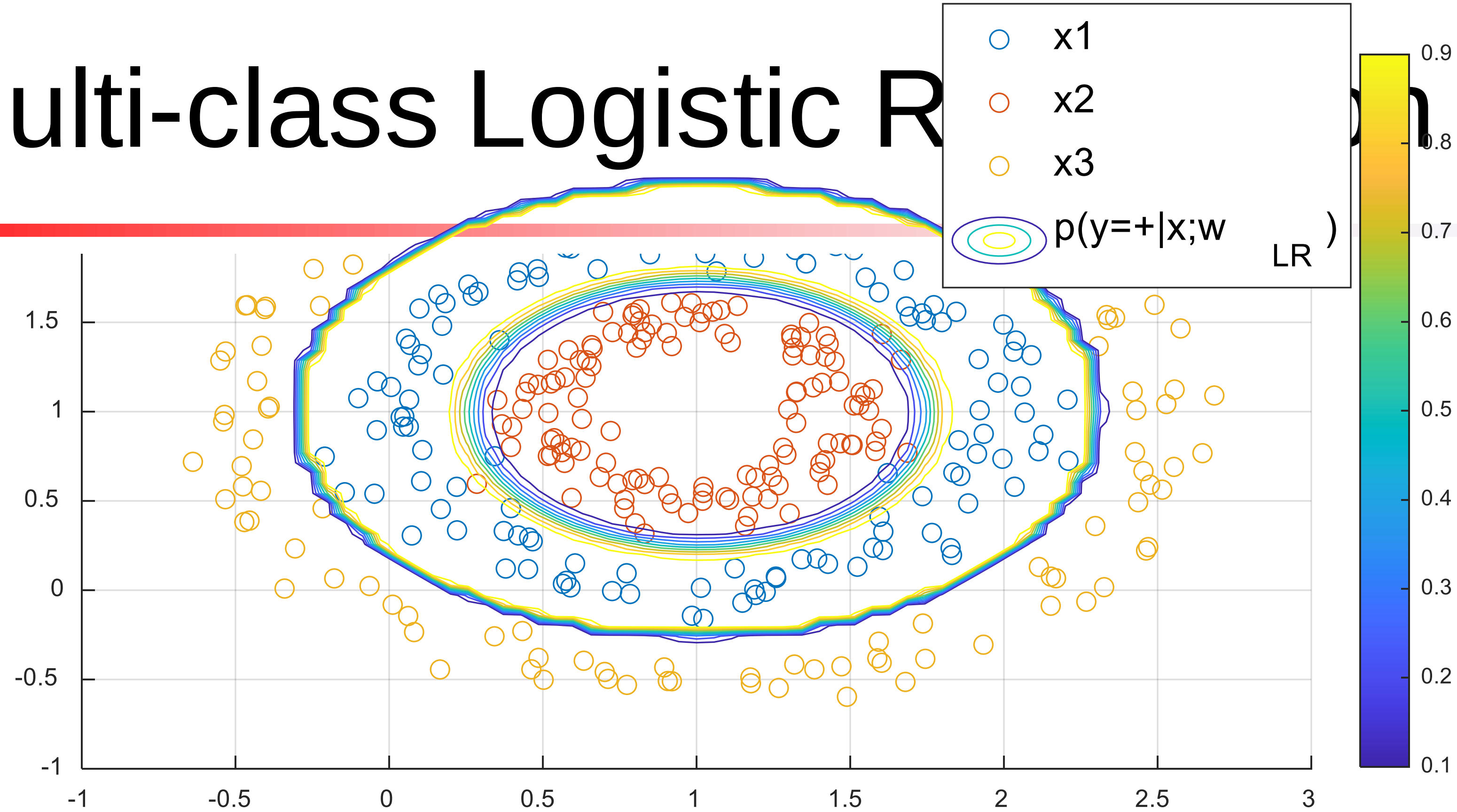
Previous Lecture

- 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 R n



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
$$\operatorname{argmax}_{\boldsymbol{w}} \sum_{i \in D} \log \sigma(f(\boldsymbol{x}_i; \boldsymbol{w}) \cdot y_i)$$
- 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 $f(\boldsymbol{x};\boldsymbol{w}) := \langle \boldsymbol{w}', \boldsymbol{x} \rangle + w_0$, 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 is the decision function given by a binary logistic regression? (hint: more than one)

- 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 $\boldsymbol{w}^*$