Special Lecture,
Pattern Information Processing

25 June, 2014

Learning under Class-Prior Change

Christo du Plessis

Department of Computer Science
Tokyo Institute of Technology

Outline

1. **Motivating Example**
2. Classification and Risk
3. Class-prior Change
4. Class-prior Change Mitigation
5. Class-prior Change Correction
6. HomeWork

Motivating Example (1)

- A certain medical test for a rare disease has a high accuracy:
  - If the disease is present, the test gives a positive result 90% of the time
  - If the disease is not present, the test gives a negative result 90% of the time

- The disease is quite rare and *only 5% of the population has the disease*
- How likely is the disease if the test result is positive?
  - Around **0.9**, **0.5**, or **0.3**?

Motivating Example (2)

- Frequency of the disease in the population

$$P(A) = 5\% \quad P(\bar{A}) = 1 - P(A) = 95\%$$

A	Disease occurs
$\bar{A}$	Disease does not occur

- Frequency of a positive result when the disease is present:

$$P(B|A) = 90\%$$

B	Test is positive
$\bar{B}$	Test is negative

- Frequency of a negative result when disease is not present

$$P(\bar{B}|\bar{A}) = 90\% \quad P(B|\bar{A}) = 10\%$$

- Frequency that the disease occurs when the test gives a positive answer: $P(A|B)$ **0.9**, **0.5**, or **0.3**?

Motivating Example (3)

- This can be computed with Bayes' rule

$$\overset{\text{Posterior}}{\boxed{P(A|B)}} = \frac{P(B|A)\overset{\text{Prior}}{\boxed{P(A)}}}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}$$

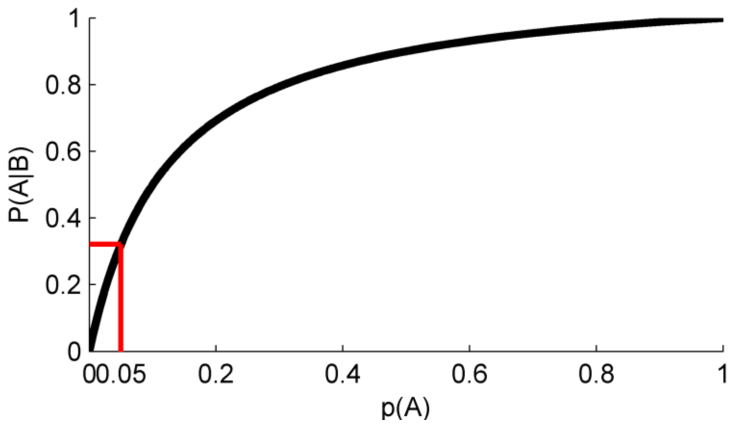
Rev. Thomas Bayes, English statistician and minister

- Substituting the values in the previous slide gives

$$P(A|B) = \frac{0.9 \times 0.05}{0.9 \times 0.05 + 0.1 \times 0.95}$$
$$= 32.13\%$$

Motivating Example (4)

- The result is counterintuitive: much lower than commonly expected
- This is due to the low class prior



$$P(A) = 5\%$$

- **Conclusion:** When doing inference, it is important to take into account the effect of the **class prior**!
- In this lecture, we will discuss the effect of the class prior on classification

Outline

1. Motivating Example
2. **Classification and Risk**
   - **Risk minimization for Classification**
   - Risk and Class-prior
3. Class-prior Change
4. Class-prior Change Mitigation
5. Class-prior Change Correction
6. Homework

Classification

- Training data:

$$\mathcal{X}_{\mathrm{tr}} := \{\boldsymbol{x}, y\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, y)$$

$\boldsymbol{x}$ Feature

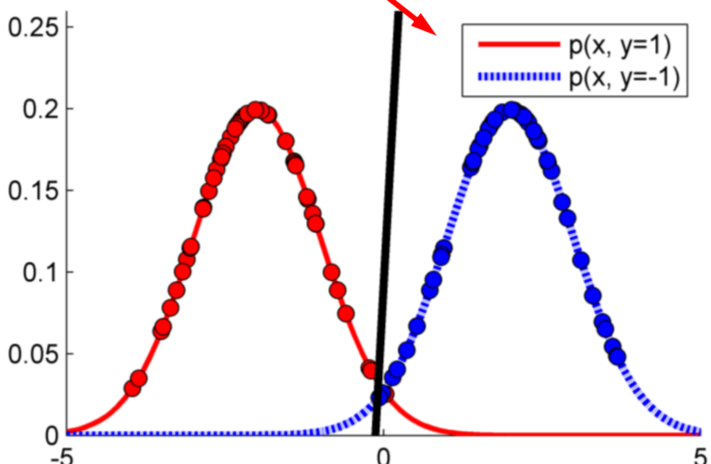
$y (\in \{-1, 1\})$ Class label

i.i.d: Independently and identically distributed

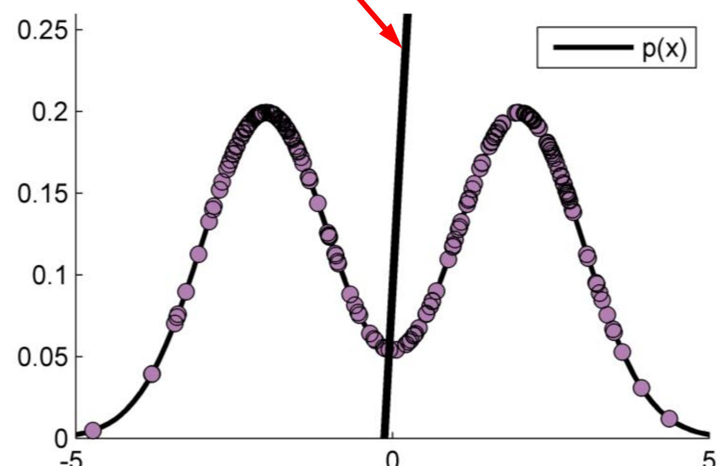
- **Goal:** Learn a rule to classify the labeled and unlabeled samples

Decision boundary learned from labeled samples

Learned decision boundary applied on unlabeled samples



Labeled training data from $p(\boldsymbol{x}, y)$

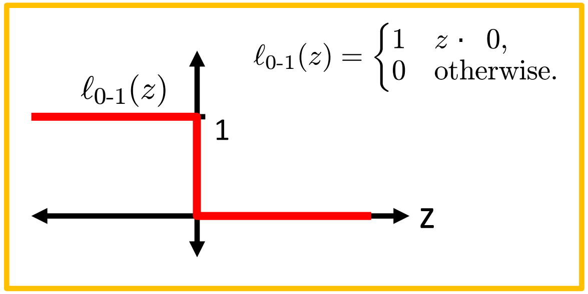


Unlabeled test data from $p(\boldsymbol{x})$

- According to what criterion should the decision boundary be selected?

Risk and Classification (1)

- $f(\boldsymbol{x})$ is a decision boundary: $f(\boldsymbol{x}) \geq 0$: Class 1; $f(\boldsymbol{x}) < 0$: Class -1



- Risk defined as

Misclassification cost **Class prior**

$$R(f) = c_{+}p(y=1)R_1(f) + c_{-1}\left[1 - p(y=1)\right]R_{-1}(f)$$

f: Discriminant

$$R_1(f) = \int \ell_{0\text{-}1}(f(\boldsymbol{x}))p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x}$$

False Negative Rate

$$R_{-1}(f) = \int \ell_{0\text{-}1}(-f(\boldsymbol{x}))p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}$$

False Positive Rate

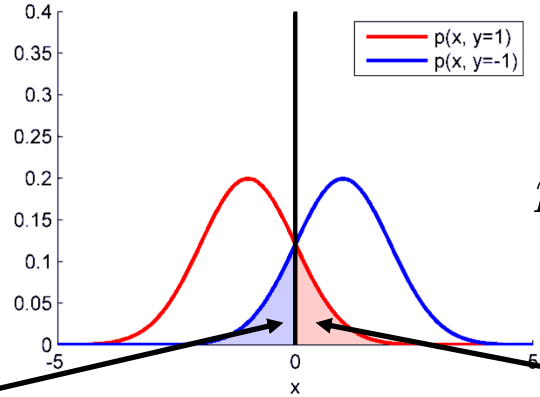
$$\ell_{0\text{-}1}(f(\boldsymbol{x})) = \begin{cases} 1 & f(\boldsymbol{x}) < 0, \\ 0 & f(\boldsymbol{x}) \geq 0. \end{cases}$$

$$\ell_{0\text{-}1}(-f(\boldsymbol{x})) = \begin{cases} 0 & f(\boldsymbol{x}) < 0, \\ 1 & f(\boldsymbol{x}) \geq 0. \end{cases}$$

Risk and Classification (2)

$$R(f) = p(y=1)R_1(f) + [1 - p(y=1)]\, R_{-1}(f)$$

$$R_1(f) = \int \ell_{0\text{-}1}(f(\boldsymbol{x}))p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x} \qquad R_{-1}(f) = \int \ell_{0\text{-}1}(-f(\boldsymbol{x}))p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}$$



$$p(\boldsymbol{x}, y=1) = p(\boldsymbol{x}|y=1)p(y=1)$$

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

$$[1 - p(y=1)] \int_{f(x)>0} p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x} \qquad p(y=1) \int_{f(x)<0} p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x}$$

- Decision boundary should minimize the risk

$$f^* = \arg\min_f R(f) \qquad R^* = R(f^*)$$

- When $c_+ = c_- = 1$, risk is the misclassification rate
- What is the optimal f^* that minimizes the risk?

Optimal classifier

- When $c_+ = c_- = 1$, the optimal discriminant is

$$f(\boldsymbol{x}) = \mathrm{sign}\left[p(y=1|\boldsymbol{x}) - p(y=-1|\boldsymbol{x})\right]$$

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)p(y)}{\sum_{y'} p(\boldsymbol{x}|y')p(y')}$$

$p(y)$	Class prior
$p(\boldsymbol{x}\|y)$	Class-conditional density
$p(y\|\boldsymbol{x})$	Posterior

- $f(x)$ is the Bayes-optimal classifier and $R^* = R(f^*)$ is the Bayes-optimal Risk

Outline

1. Motivating Example
2. **Classification and Risk**
   - Risk minimization for Classification
   - **Risk and Class-prior**
3. Class-prior Change
4. Class-prior Change Mitigation
5. Class-prior Change Correction
6. HomeWork

Bayes Risk vs. Class prior

- Recall, Bayes Risk is

We will use the symbol π to denote a class prior $p(y=1)$ from now on

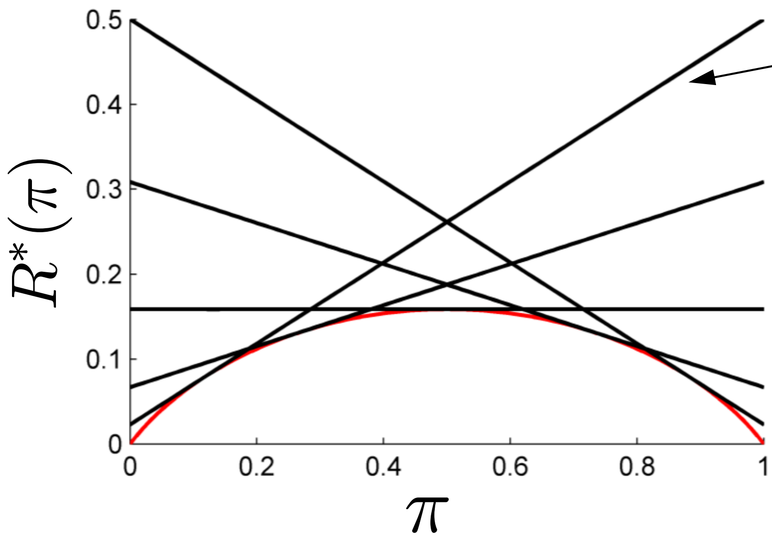
$$R^*(\pi) = \min_f \pi R_1(f) + (1-\pi)R_{-1}(f)$$

$$\boxed{R_1(f) = \int \ell_{0\text{-}1}(f(\boldsymbol{x})p(\boldsymbol{x}|y=1)\mathrm{d}\boldsymbol{x}}$$

False Negative Rate

$$\boxed{R_{-1}(f) = \int \ell_{0\text{-}1}(-f(\boldsymbol{x})p(\boldsymbol{x}|y=-1)\mathrm{d}\boldsymbol{x}}$$

False Positive Rate

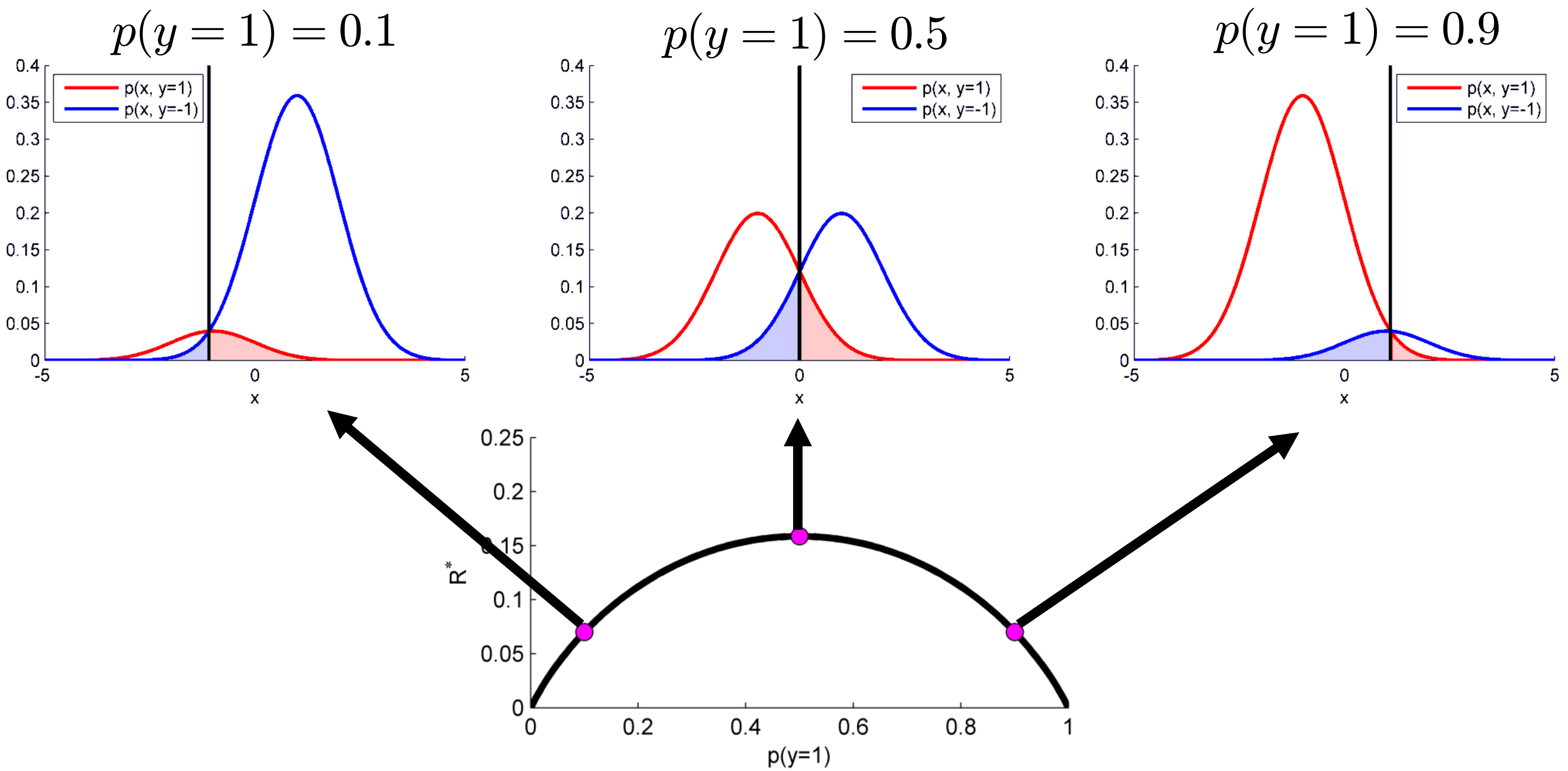


$\pi R_1(f_0) + (1-\pi)R_{-1}(f_0)$

Function is *concave* w.r.t. π:

- Minimum of linear functions

Example



- The decision boundary changes when the class prior changes
- Bayes risk is dependent on the class prior

Conclusion: Section 2

- The misclassification rate is a weighted combination of the false negative and false positive rate
  - Weighted by the class priors
- The classifier that minimizes the risk is

$$f(\boldsymbol{x}) = \operatorname{sign}\left[p(y=1|\boldsymbol{x}) - p(y=-1|\boldsymbol{x})\right]$$

- The optimal risk is a *concave function* of the class prior

Outline

1. Motivating Example
2. Classification and Risk
3. **Class-prior Change**
   - **Class-prior Change**
   - Causes of prior Change
     - Dataset shift
     - Selection Bias
   - Class-prior change and Risk
4. Class-prior Change Mitigation
5. Class-prior Change Correction
6. Homework

Class-prior Change

- Class prior between the training and test data differ:

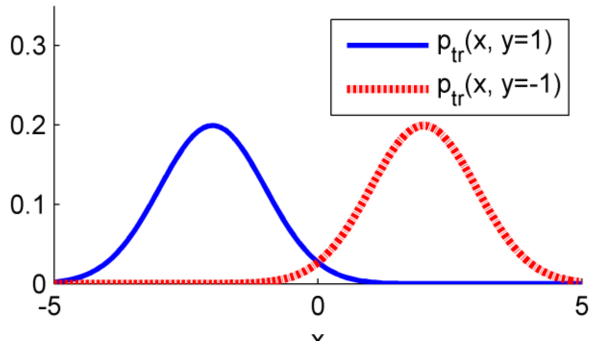
$$p_{\mathrm{te}}(\boldsymbol{x}, y) = p(\boldsymbol{x}|y)p_{\mathrm{te}}(y) \quad p_{\mathrm{tr}}(\boldsymbol{x}, y) = p(\boldsymbol{x}|y)p_{\mathrm{tr}}(y)$$

$$p_{\mathrm{tr}}(y) \neq p_{\mathrm{te}}(y)$$ Class priors differ

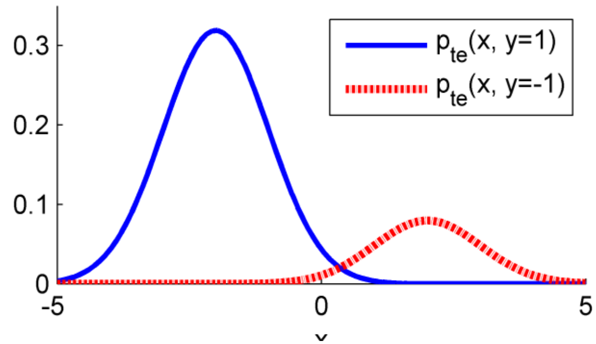
$$p(\boldsymbol{x}|y)$$ Same class-conditional density

- Example:

Training distribution



Test distribution



Outline

1. Motivating Example
2. Classification and Risk
3. **Class-prior Change**
   - Class-prior Change
   - **Causes of Class-prior Change**
     - **Dataset shift**
     - **Selection Bias**
   - Class-prior Change and Risk
4. Class-prior Change Mitigation
5. Class-prior Change Correction
6. Homework

Why may the dataset change?

- Dataset shift
  - Natural change in the dataset between training and test
  - Example: Face images selected in a laboratory compared to the real world

Training dataset:

(Olivetti dataset)

Class balance:
Male: 18/20
Female: 2/20

- Selection bias (next slide)

Selection bias

- Samples drawn for the test dataset may be drawn in a biased way
- Selection bias model

$$(\boldsymbol{x}, y, s) \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, y, s)$$

$\boldsymbol{x}$	Feature
$y(\in \{-1, 1\})$	Class label
$s(\in \{0, 1\})$	Selection of samples

- When $s = 1$, the sample is in the test set, when $s = 0$, the sample is not in the test set
- Training distribution: $p_{\mathrm{tr}}(\boldsymbol{x}, y) = p(\boldsymbol{x}, y)$
- Test distribution: $p_{\mathrm{te}}(\boldsymbol{x}, y) = p(\boldsymbol{x}, y | s = 1)$

Selection bias (2)

- Three possibilities:
  1. **No selection bias**
     s is independent of x and y
     $$p(s=1|x,y)=p(s=1)$$
     No selection bias
  2. **Covariate shift**
     s is independent of y given x
     $$p(s=1|x,y)=p(s=1|x)$$
  3. **Class-prior change**
     s is independent of x given y
     $$p(s=1|x,y)=p(s=1|y)$$

Covariate shift

- $p(s=1|\boldsymbol{x},y) = p(s=1|\boldsymbol{x})$ implies that

$$p_{\mathrm{tr}}(y|\boldsymbol{x}) = p_{\mathrm{te}}(y|\boldsymbol{x})$$

- Proof:

$$\begin{aligned} p_{\mathrm{te}}(y|\boldsymbol{x}) = p(y|\boldsymbol{x}, s=1) &= \frac{p(\boldsymbol{x}, y, s=1)}{p(\boldsymbol{x}, s=1)} \\ &= \frac{p(s=1|\boldsymbol{x}, y)p(\boldsymbol{x}, y)}{p(\boldsymbol{x}, s=1)} \\ &= \frac{p(s=1|\boldsymbol{x})p(\boldsymbol{x}, y)}{p(\boldsymbol{x}, s=1)} \\ &= \frac{p(s=1|\boldsymbol{x})}{p(s=1|\boldsymbol{x})} p(y|\boldsymbol{x}) = p_{\mathrm{tr}}(y|\boldsymbol{x}) \end{aligned}$$

$$\boxed{\begin{aligned} p_{\mathrm{te}}(\boldsymbol{x}, y) &= p(\boldsymbol{x}, y|s=1) \\ p_{\mathrm{tr}}(\boldsymbol{x}, y) &= p(\boldsymbol{x}, y) \end{aligned}}$$

- Covariate shift occurs in practice!
- Can be corrected for in the semi-supervised setup
- See **lecture 13**

Class-prior Change

- $p(s = 1|\boldsymbol{x}, y) = p(s = 1|\boldsymbol{y})$ implies that

$$p_{\mathrm{tr}}(\boldsymbol{x}|y) = p_{\mathrm{te}}(\boldsymbol{x}|y)$$

- Proof:

$$\boxed{\begin{aligned} p_{\mathrm{te}}(\boldsymbol{x}, y) &= p(\boldsymbol{x}, y|s = 1) \\ p_{\mathrm{tr}}(\boldsymbol{x}, y) &= p(\boldsymbol{x}, y) \end{aligned}}$$

$$\begin{aligned} p_{\mathrm{te}}(\boldsymbol{x}|y) = p(\boldsymbol{x}|y, s = 1) &= \frac{p(\boldsymbol{x}, y, s = 1)}{p(y, s = 1)} \\ &= \frac{p(s = 1|\boldsymbol{x}, y)p(\boldsymbol{x}, y)}{p(s = 1|y)p(y)} \\ &= \frac{p(s = 1|y)}{p(s = 1|y)}\frac{p(\boldsymbol{x}, y)}{p(y)} = p(\boldsymbol{x}|y) \end{aligned}$$

- Class-prior change may be due to selection bias
- We discuss methods to mitigate the effect of class-prior change in the supervised setup
- We discuss a correction for class-prior change in the semi-supervised setup

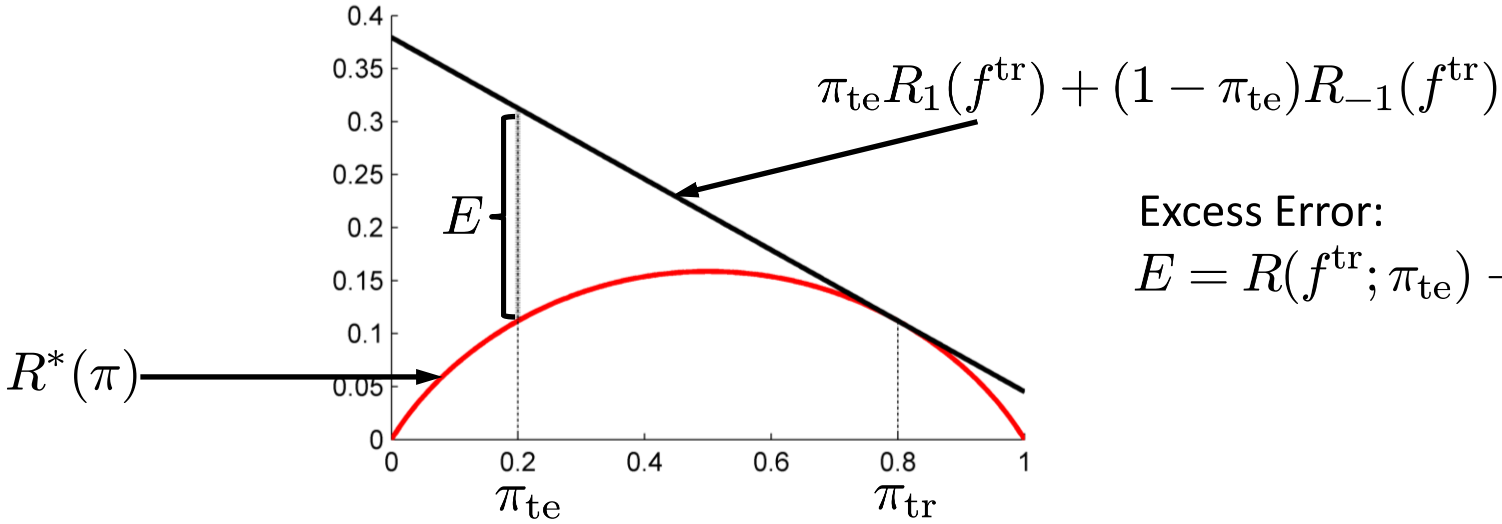
Outline

1. Motivating Example
2. Classification and Risk
3. **Class-prior Change**
   - Class-prior Change
   - Causes of Class-prior Change
     - Dataset shift
     - Selection Bias
   - **Class-prior Change and Risk**
4. Class-prior Change Mitigation
5. Class-prior Change Correction
6. Homework

Effect of Class-prior Change on Risk

- Training: $f_{\mathrm{tr}} = \arg\min_{f} R\,(f\,; \pi_{\mathrm{tr}})$ $R_1(f_{\mathrm{tr}})$ $R_{-1}(f_{\mathrm{tr}})$
- At point π_{tr} this is the same as $R^*(\pi_{\mathrm{tr}})$
- When f_{tr} is applied to a dataset with class prior π_{te} the error is

$$\pi_{\mathrm{te}} R_1(f^{\mathrm{tr}}) + (1 - \pi_{\mathrm{te}}) R_{-1}(f^{\mathrm{tr}}) \geq \min_{f} R(\pi_{\mathrm{te}})$$



Excess Error:

$E = R(f^{\mathrm{tr}}; \pi_{\mathrm{te}}) - R^*(\pi_{\mathrm{te}})$

Conclusion: Section 3

- *Class-prior change* may occur between the training and test data:

$$p_{\mathrm{te}}(\boldsymbol{x}, y) = p(\boldsymbol{x}|y)p_{\mathrm{te}}(y) \quad p_{\mathrm{tr}}(\boldsymbol{x}, y) = p(\boldsymbol{x}|y)p_{\mathrm{tr}}(y)$$

$$p_{\mathrm{tr}}(y) \neq p_{\mathrm{te}}(y)$$

- This may be due to *dataset shift* or *sample selection bias*
- When a classifier is selected according to π_{tr} and applied on a dataset with π_{te}, the Risk is linear and tangent to the optimal risk curve at π_{tr}

Outline

1. Motivating Example
2. Classification and Risk
3. Class-prior Change
4. **Class-prior Change Mitigation**
   - **The minimax criterion**
5. Class-prior Change Correction
6. Homework

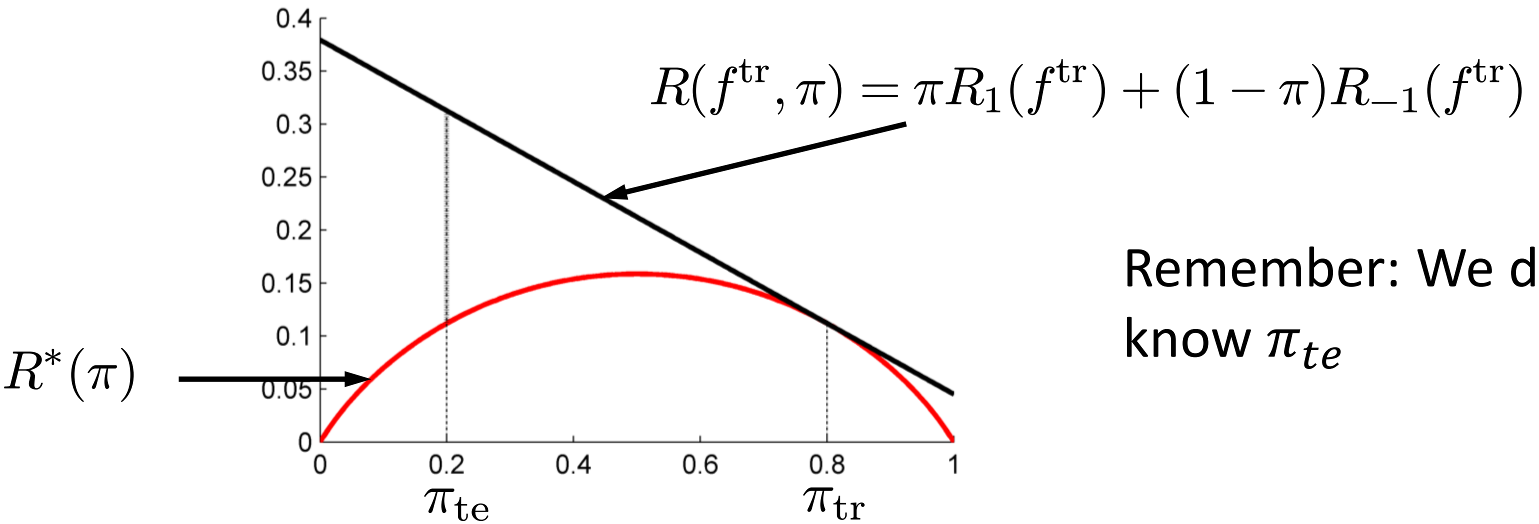
Class-prior change mitigation

- Class-prior change can have an adverse effect on the classification accuracy
- In practice, the test class prior π_{te} is unknown
  - We can therefore not correct for the effect of class-prior change
- Can we mitigate the effect of class prior change?

Mitigate (definition): to make less severe

Minimax Criterion (1)

- Recall this figure:

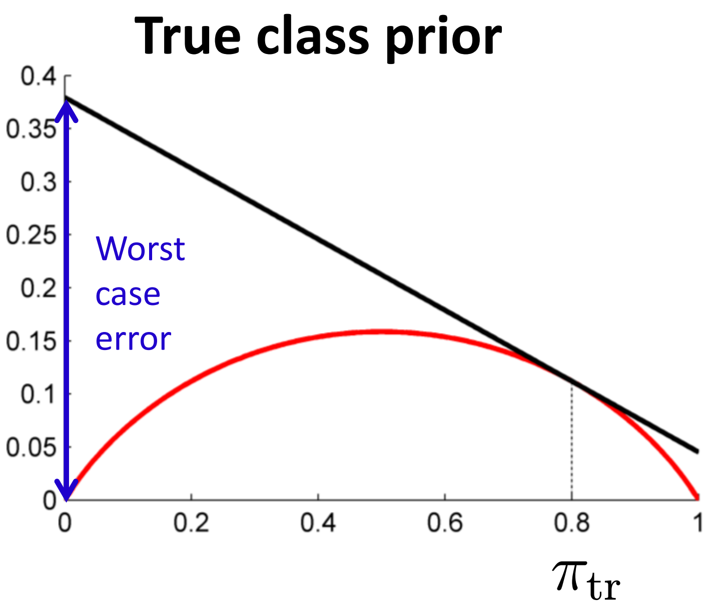


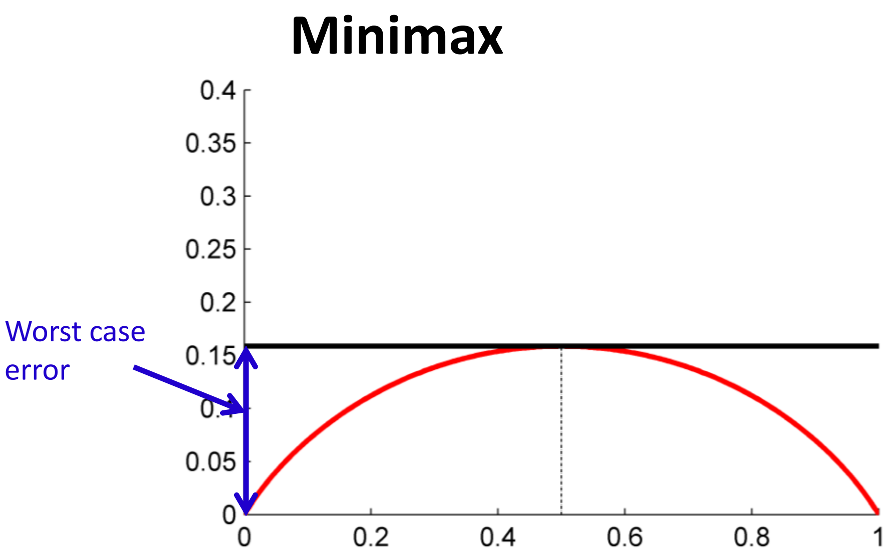
Remember: We do not know π_{te}

- The black line is the misclassification rate according to the new class prior when trained with the old class prior
- This line is always tangent to the optimal risk $R^*(\pi)$

Minimax Criterion (2)

- Why not select this line so that it does not change w.r.t. the new class prior?
- In other words, the tangent should be 0
- Since $R^*(\pi)$ is concave, this would occur at the maximum





Further Reading

- The minimax criterion is described in
  - "*Pattern Classification*", 2^{nd} Edition (Richard O. Duda, Peter E. Hart, David G. Stork), p.g. 26.
    - *(Ookayama Main Lib. B1F - Books 548.13/D)*
  - "*Detection, estimation, and linear modulation theory*" (Van Trees, Harry L.) *1968*
    - Ookayama Main Lib. B1F ; Compact Shelving - Y000998
  - The minimax criterion is discussed, and an extention introduced in "*Minimax Regret Classifier for Imprecise Class Distributions*" (Alaiz-Rodríguez, Rocío, Alicia Guerrero-Curieses, and Jesús Cid-Sueiro)
    - http://jmlr.org/papers/volume8/alaiz-rodriguez07a/alaiz-rodriguez07a.pdf

Outline

1. Motivating Example
2. Classification and Risk
3. Class-prior Change
4. Class-prior Change Mitigation
5. **Class-prior Change Correction**
   - **Classifier reweighting**
   - Class-prior Estimation in the Semi-Supervised Setup
   - Class-prior Estimation via Distribution Matching
   - Distribution matching via L_2-distance minimization
   - Direct L_2-distance Estimation
   - Example
6. Homework

Correction for Class-prior change

- Recall that a cost-sensitive classifier minimizes

$$R(f) = c_{+}\pi_{\mathrm{tr}}R_{1}(f) + c_{-1}\left[1 - \pi_{\mathrm{tr}}\right]R_{-1}(f)$$

False negative rate **False positive rate**

- Misclassification rate according to π_{te} can be obtained by weights:

$$c_{+} = \frac{\pi_{\mathrm{te}}}{\pi_{\mathrm{tr}}} \qquad c_{-} = \frac{1 - \pi_{\mathrm{te}}}{1 - \pi_{\mathrm{tr}}}$$

- Libraries such as libSVM allows specification of cost
- **Problem:** Test class prior π_{te} is often unknown

Outline

1. Motivating Example
2. Classification and Risk
3. Class-prior Change
4. Class-prior Change Mitigation
5. **Class-prior Change Correction**
   - Classifier reweighting
   - **Class-prior Estimation in the Semi-Supervised Setup**
   - Class-prior Estimation via Distribution Matching
   - Distribution matching via L_2-distance minimization
   - Direct L_2-distance Estimation
   - Example
6. Homework

Semi-supervised setup

- In many situations, unlabeled data in addition to labeled data is available

$$\mathcal{X}_{\mathrm{tr}} := \{\boldsymbol{x}, y\}_{i=1}^{n} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{tr}}(\boldsymbol{x}, y) \qquad \mathcal{X}_{\mathrm{te}} := \{\boldsymbol{x}_i\}_{i=1}^{n'} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{te}}(\boldsymbol{x})$$

- In the class-prior change assumption, the two distributions shares a class-conditional density:

$$p_{\mathrm{tr}}(\boldsymbol{x}, y) = \underbrace{\boxed{p(\boldsymbol{x}|y)}}_{\text{Shared}} p_{\mathrm{tr}}(y) \qquad p_{\mathrm{te}}(\boldsymbol{x}, y) = \underbrace{\boxed{p(\boldsymbol{x}|y)}}_{\text{Shared}} p_{\mathrm{te}}(y)$$

- We wish to estimate the class prior of the unlabeled dataset $p_{\mathrm{te}}(y)$
- This is difficult, because no labeled samples are available

Outline

1. Motivating Example
2. Classification and Risk
3. Class-prior Change
4. Class-prior Change Mitigation
5. **Class-prior Change Correction**
   - Classifier reweighting
   - Class-prior Estimation in the Semi-Supervised Setup
   - **Class-prior Estimation via Distribution Matching**
   - Distribution matching via L_2-distance minimization
   - Direct L_2-distance Estimation
   - Example
6. Homework

Distribution matching framework

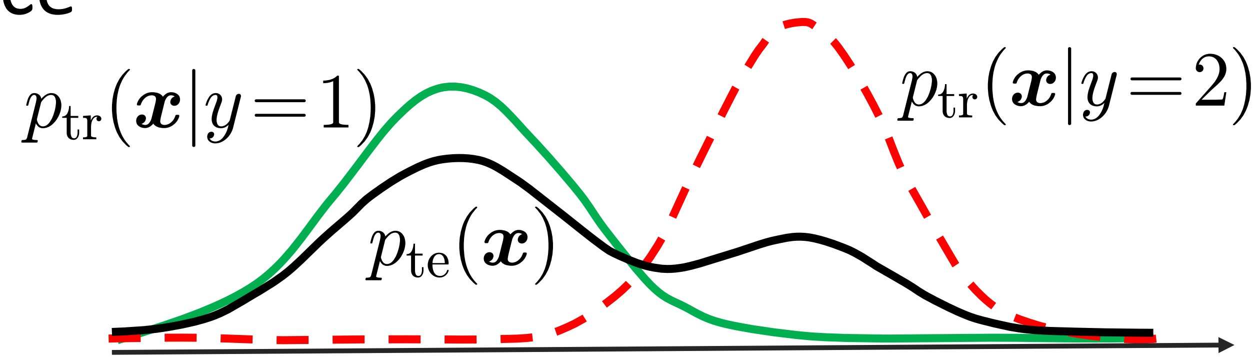
- Lets model the test input distribution $p_{\mathrm{te}}(\boldsymbol{x})$ in terms of:
  - The training class-conditional distribution $p_{\mathrm{tr}}(\boldsymbol{x}|y)$
  - The test class priors $\pi_y = p_{\mathrm{te}}(y)$

$$q_{\mathrm{te}}(\boldsymbol{x}) = \sum_{y=1}^{c} \pi_y p_{\mathrm{tr}}(\boldsymbol{x}|y)$$

Sicne $p_{tr}(x|y) = p_{te}(x|y)$:

$$p_{\mathrm{te}}(\boldsymbol{x}) = \sum_{y=1}^{c} p_{\mathrm{tr}}(\boldsymbol{x}|y) p_{\mathrm{te}}(y)$$

- **Problem:** Match $q_{\mathrm{te}}(\boldsymbol{x})$ to $p_{\mathrm{te}}(\boldsymbol{x})$ under some divergence



Outline

1. Motivating Example
2. Classification and Risk
3. Class-prior Change
4. Class-prior Change Mitigation
5. **Class-prior Change Correction**
   - Classifier reweighting
   - Class-prior Estimation in the Semi-Supervised Setup
   - Class-prior Estimation via Distribution Matching
   - **Distribution matching via L_2-distance minimization**
   - Direct L_2-distance Estimation
   - Example
6. Homework

Distribution Matching via L_2-distance Minimization

- The similarity between two distributions can be measured by the L_2-distance:

$$L_2\left(p_{\mathrm{te}}, q_{\mathrm{te}}\right) = \frac{1}{2}\int \left[p_{\mathrm{te}}(\boldsymbol{x}) - q_{\mathrm{te}}(\boldsymbol{x})\right]^2 \mathrm{d}\boldsymbol{x}$$

- The class prior can therefore be selected as

$$(\pi_1, \ldots, \pi_c) = \operatorname*{arg\,min}_{\pi} L_2\left(p_{\mathrm{te}}(\boldsymbol{x}), q_{\mathrm{te}}(\boldsymbol{x})\right)$$

- The L_2 distance can be estimated by first estimating the densities $p_{te}(x)$ and $q_{te}(x)$
  - Not good since density estimation is a difficult problem

Outline

1. Motivating Example
2. Classification and Risk
3. Class-prior Change
4. Class-prior Change Mitigation
5. **Class-prior Change Correction**
    - Classifier reweighting
    - Class-prior Estimation in the Semi-Supervised Setup
    - Class-prior Estimation via Distribution Matching
    - Distribution matching via L_2-distance minimization
    - **Direct L_2-distance Estimation**
    - Example
6. Homework

Direct L_2 estimation

- Expectations can be estimated via sample averages:

$$\int f(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \mathbb{E}_p\left[f(\boldsymbol{x})\right] \approx \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) \qquad \{\boldsymbol{x}_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})$$

- Can we estimate the L_2-distance in terms of sample averages?
- It is possible by obtaining a lower-bound that is linear in the densities

Direct L_2 estimation

- We use the following inequality:

$$\frac{1}{2}(t-y)^2 \geq 0, \Rightarrow \frac{1}{2}t^2 \geq ty - \frac{1}{2}y^2$$

- Applying this pointwise gives

$w(\boldsymbol{x})$ role of y

$$\frac{1}{2}\left[p_{\mathrm{te}}(\boldsymbol{x}) - q_{\mathrm{te}}(\boldsymbol{x})\right]^2 \geq w(\boldsymbol{x})\left[p_{\mathrm{te}}(\boldsymbol{x}) - q_{\mathrm{te}}(\boldsymbol{x})\right] - \frac{1}{2}w(\boldsymbol{x})^2$$

- Integrating and selecting the tightest lower-bound gives

$$\frac{1}{2}\int \left[p_{\mathrm{te}}(\boldsymbol{x}) - q_{\mathrm{te}}(\boldsymbol{x})\right]^2 \mathrm{d}\boldsymbol{x}$$

$$\geq \sup_{w} \int w(\boldsymbol{x})\left[p_{\mathrm{te}}(\boldsymbol{x}) - q_{\mathrm{te}}(\boldsymbol{x})\right] \mathrm{d}\boldsymbol{x} - \frac{1}{2}\int w(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x}$$

Direct L_2 estimation

- This lower-bound can then be estimated via sample averages
- Lets model $w(x)$ with a linear model

$$w(\boldsymbol{x}) = \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}) \qquad \varphi_\ell(\boldsymbol{x}) = \exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{x} - \boldsymbol{c}_\ell\right\|^2\right)$$

- The L_2-distance lower bound can be written in terms of expectations

$$L_2(p_{\mathrm{te}}, q_{\mathrm{te}}) \geq \sup_w \mathbb{E}_{p_{\mathrm{te}}}\left[w(\boldsymbol{x})\right] - \sum_{c=1}^{n_c} \mathbb{E}_{\pi_y p_{\mathrm{te}}}\left[w(\boldsymbol{x})\right] - \frac{1}{2}\int w(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x}$$

Direct L_2 estimation

$$L_2(p_{\mathrm{te}}, q_{\mathrm{te}}) \geq \sup_w \mathbb{E}_{p_{\mathrm{te}}}\left[w(\boldsymbol{x})\right] - \sum_{c=1}^{n_c} \mathbb{E}_{\pi_y p_{\mathrm{te}}}\left[w(\boldsymbol{x})\right] - \frac{1}{2}\int w(\boldsymbol{x})^2 \mathrm{d}\boldsymbol{x}$$

- The expectations can be estimated via sample averages:

$$\widehat{\boldsymbol{h}}_{te} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{x}_i) \qquad \widehat{\boldsymbol{h}}_c = \frac{1}{n_c}\sum_{i=1, y_i=c}^{n} \boldsymbol{\varphi}(\boldsymbol{x}_i) \qquad \boldsymbol{H} = \int \boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^\top \mathrm{d}\boldsymbol{x}$$

$$\boldsymbol{\varphi}(\boldsymbol{x}) = \left[\varphi_1(\boldsymbol{x}) \quad \varphi_2(\boldsymbol{x}) \quad \ldots \quad \varphi_\ell(\boldsymbol{x})\right]$$

- Which gives an objective function of:

$$\widehat{L}_2(\{\pi_y\}_{y=1}^{c}) \approx \max_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^\top \widehat{\boldsymbol{h}}_{\mathrm{te}} - \sum_{y=1}^{c} \theta_y \boldsymbol{\alpha}^\top \widehat{\boldsymbol{h}}_y - \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{H}\boldsymbol{\alpha}$$

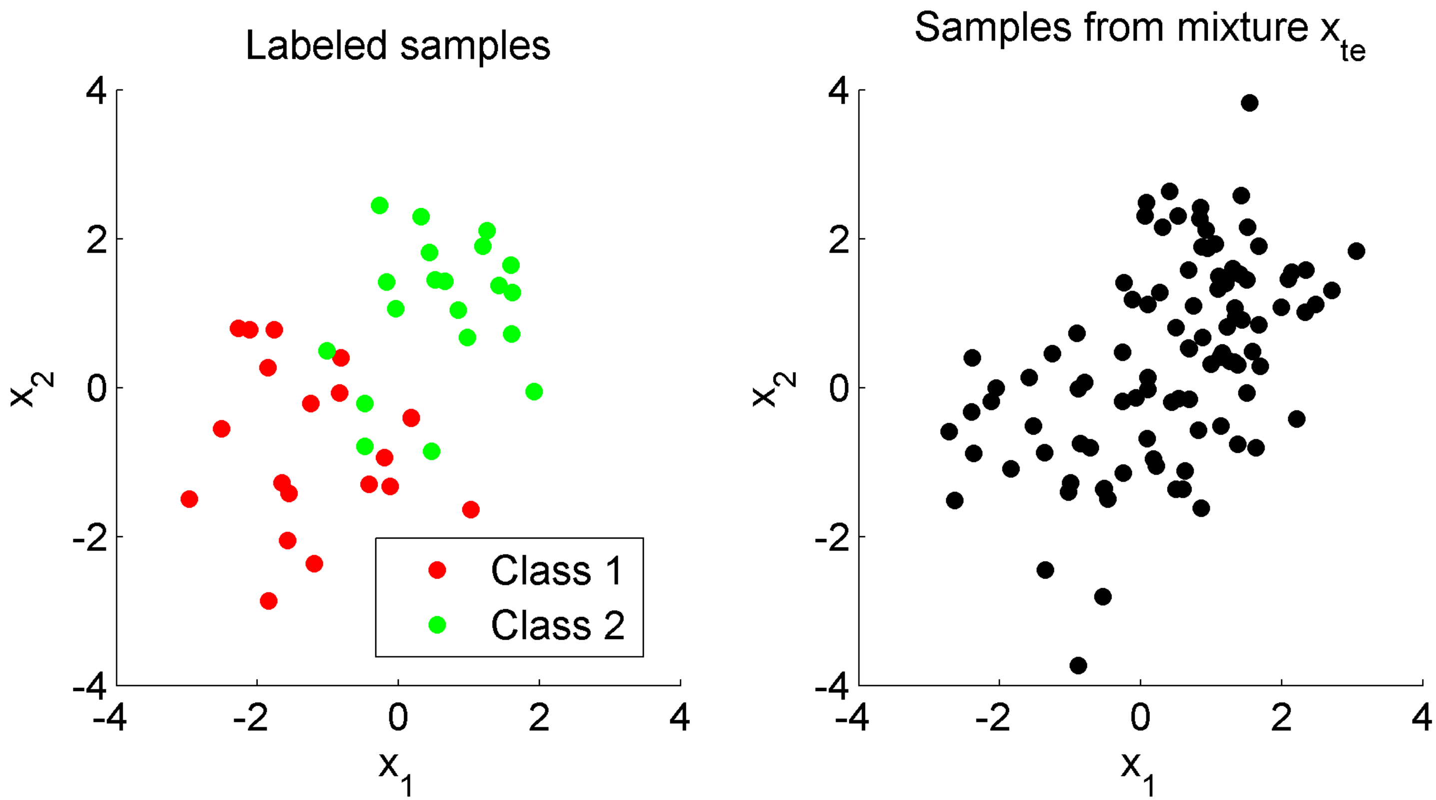
Direct L_2 estimation

- Minimizing the L_2 distance estimate w.r.t. $\{\pi_y\}_{y=1}^{c}$ gives an estimate of the class prior
- This can then be used to reweight a classifier

Outline

Example

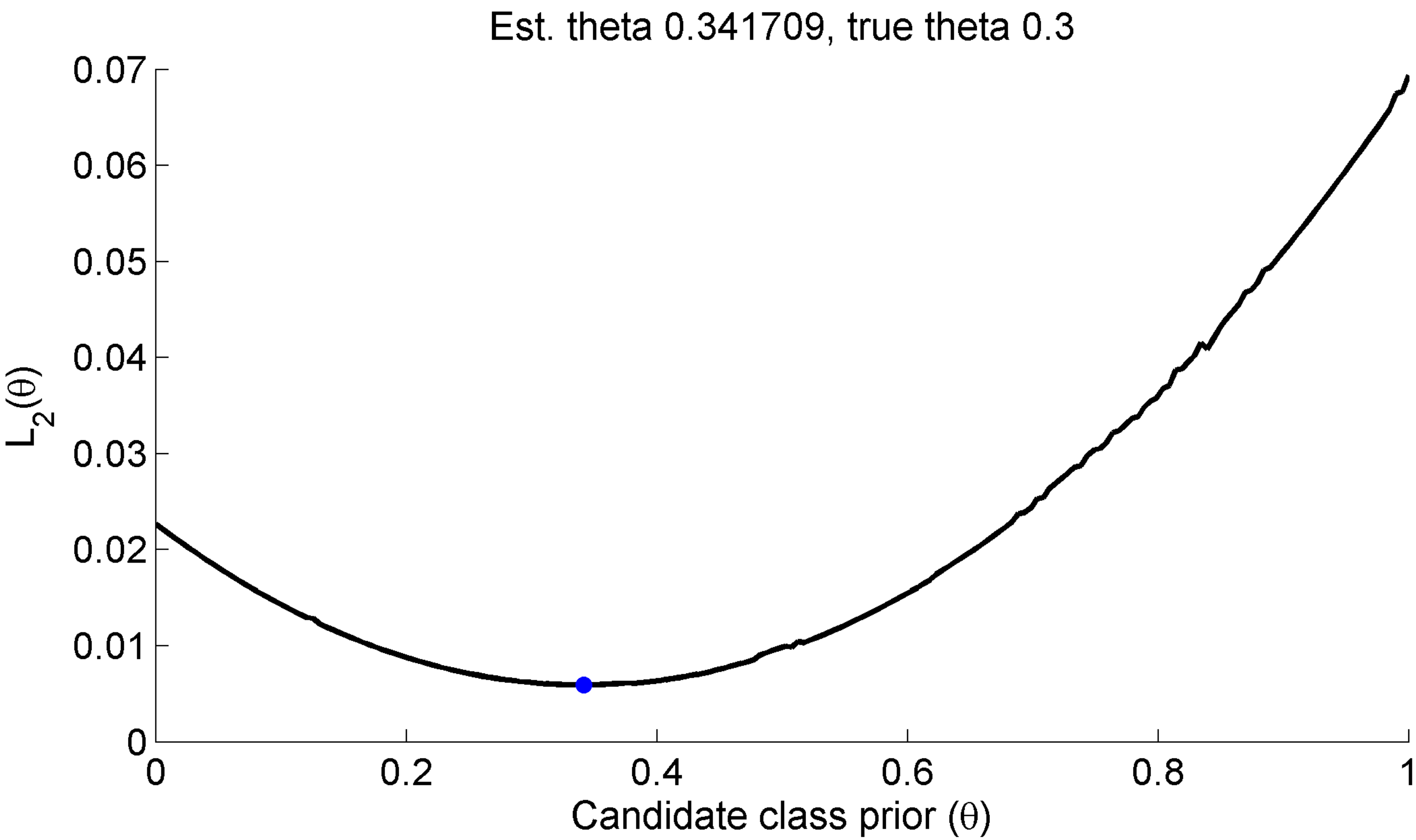
- Samples from two Gaussians with different means:



- The true class prior is $p_{te}(y = 1) = 0.3$

Example (2)

- The L_2-distance estimated from samples is given below:



- Minimum of L_2 distance is near the true class prior
- Difference is due to estimation from a small set l

Conclusion: Section 5

- By *reweighting* the risk, a classifier can be trained
- The *reweighting factor* depends on the *unknown class prior*
- In a *semi-supervised setup*, the *unknown class prior can be estimated*
- Estimation is possible by matching a model of the test input density to the true test input density

Further Reading

- "Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure" (Saerens, M., Latinne, P., and Decaestecker, C.)
  - Neurocomputation 14 (2002)
  - Introduced estimation of the class prior for re-adjustment of the classifier
- "Semi-supervised learning of class balance under class-prior change by distribution matching." (du Plessis, M. C. & Sugiyama, M.)
  - Estimation of the class prior via Pearson divergence matching
- "Density-difference estimation" (Sugiyama, M., Suzuki, T., Kanamori, T., du Plessis, M. C., Liu, S., & Takeuchi, I.)
  - Estimation of the class prior via L_2-distance estimation (discussed here)

Outline

1. Motivating Example
2. Classification and Risk
3. Class-prior Change
4. Class-prior Change Mitigation
5. Class-prior Change Correction
6. **Take-home message and homework**

Take-home Message

- The class-prior may change between the training and test phase
- In *supervised learning*, the minimax approach can be used:
  - Minimizes the worst case result
- If the test class-prior is known, the classifier can be selected by reweighting
- In *semi-supervised learning*, the test class prior can be estimated

Homework

- Please hand in your reports now!
- Homework:

Write you opinion about the special lecture today.
Directly submit the printed report to the lecturer next week.