Chapter 2
Classification: basic concepts

Esa Ollila

Department of Signal Processing and Acoustics
Aalto University, Finland

Large Scale Data Analysis / Aalto University

Training data

- We have a set of input variables (or features)

$$X = (X_1, \ldots, X_p)$$

  that are used to predict the output variable $Y \in \mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_K\}$.

- Training data

$$\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$$

  is available from the joint distribution of (X, Y).

- We encode the labels of Y with numeric values, such as

$$\mathcal{G} = \{1, \ldots, K\}$$

  in the general $K > 2$ case and

$$\mathcal{G} = \{1, 2\} \quad \text{or} \quad \mathcal{G} = \{-1, 1\}$$

  in the two-class ($K = 2$) case.

Classification task

- Classification task: problem of partitioning the input vector space into disjoint (decision) regions.
- Is tantamount to finding a discriminant rule (aka classifier) $G(\mathbf{x}) : \mathbb{R}^p \to \mathcal{G} = \{1, \ldots, K\}$, i.e., a function that takes a data vector and returns a class label.
- Discriminant rule partitions the input space $\mathbb{R}^p$ into K decision regions,

$$\Gamma_k \equiv \Gamma_k(G) = \big\{\mathbf{x} \in \mathbb{R}^p : G(\mathbf{x}) = k\big\}$$

which are mutually disjoint and exhaustive sets, i.e.,

$$\Gamma_k \cap \Gamma_j = \emptyset \;\; \forall k \neq j \quad \text{and} \quad \bigcup_{k=1}^{K} \Gamma_k = \mathbb{R}^p.$$

Classification task (cont'd)

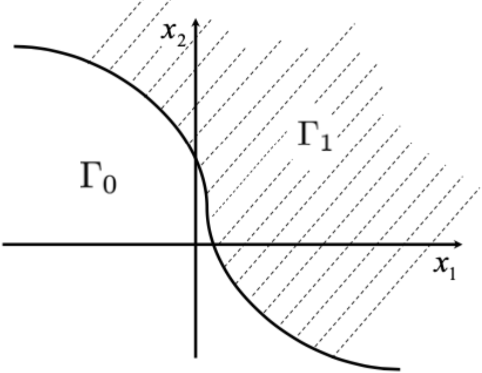
- $G(\mathbf{x})$ can often expressed as

$$G(\mathbf{x}) = \underset{k \in \{1,\dots,K\}}{\arg\max}\ G_k(\mathbf{x})$$

where $G_k(\mathbf{x}) \in \mathbb{R}$ is a discriminant score of $\mathbf{x}$ for class k, and

$$\Gamma_k = \left\{ \mathbf{x} \in \mathbb{R}^p : G_k(\mathbf{x}) > G_j(\mathbf{x}) \quad \forall k \neq j \in \{1, \dots, K\} \right\}$$

(and decide ties by random guesses).



Menu

Basic definitions

- Joint $(p+1)$-variate distribution of (Y, X) is known.
- $Y \in \{1, \ldots, K\}$ is a discrete random variable with a probability mass function (p.m.f.)

$$\pi_k = \Pr(Y = k) = \begin{cases} \text{"probability that a randomly selected} \\ \text{observation is from class } k\text{"} \end{cases}$$

- π_k-s are known a priori class probabilities ($\sum_{i=1}^K \pi_i = 1$).
- Assume $X = (X_1, \ldots, X_p)$ is a continuous p-variate random vector.
- Them it has class conditional probability density function (p.d.f.)

$$f_{X|Y}(\mathbf{x} \mid k), \quad k = 1, \ldots, K.$$

- Note also that

$$\Pr(X \in \mathcal{X} \mid Y = k) = \int_{\mathcal{X}} f_{X|Y}(\mathbf{x} \mid k) \mathrm{d}\mathbf{x}.$$

Basic definitions (cont'd)

- The a posteriori class probabilities are

$$p_k(\mathbf{x}) = \Pr(Y = k \mid X = \mathbf{x}) = \frac{f_{X|Y}(\mathbf{x} \mid k)\,\Pr(Y = k)}{f_X(\mathbf{x})}$$
$$= \frac{f_{X|Y}(\mathbf{x} \mid k)\pi_k}{\sum_{k=1}^K f_{X|Y}(\mathbf{x} \mid k)\pi_k}, \ \ k = 1, \ldots, K.$$

- Note that

$$0 \leq p_k(\mathbf{x}) \leq 1 \quad \text{and} \quad \sum_{k=1}^K p_k(\mathbf{x}) = 1.$$

- We express the log-posterior by

$$\log p_k(\mathbf{x}) = \ln f_{X|Y}(\mathbf{x} \mid k) + \ln \pi_k,$$

where we ignore constant $(-\ln f_X(\mathbf{x}))$ that does not depend on k.

Bayes risk

- The classification loss or 0-1 loss is defined as

$$L_{0/1}(y, G(\mathbf{x})) = 1_{\{y \neq G(\mathbf{x})\}}$$

  and equals 1 if the classifier G misclassifies $(\mathbf{x}, y)$, and 0 otherwise.
- The (Bayes) risk of a discriminant rule G,

$$\mathrm{r}(G) = \Pr(G(X) \neq Y) = \mathbb{E}\big[1_{\{Y \neq G(X)\}}\big],$$

  equals the probability that the rule makes an error.
- We may write it as

$$\begin{aligned} \mathrm{r}(G) &= \sum_{k \in \mathcal{G}} \mathbb{E}_X \left[1_{\{G(X) \neq k\}} \middle| Y = k\right] \pi_k \\ &= \sum_{k \in \mathcal{G}} \Pr(X \notin \Gamma_k(G) \mid Y = k) \pi_k \end{aligned}$$

- Its empirical version is called empirical risk:

$$\hat{\mathrm{r}}_N(G) = \frac{1}{N} \sum_{i=1}^{N} L_{0/1}(y_i, G(\mathbf{x}_i)) = \frac{1}{N} \sum_{i=1}^{N} 1_{\{y_i \neq G(\mathbf{x}_i)\}}$$

The Bayes classifier

Bayes classifier G^* assigns $\mathbf{x}$ to class k^* having maximum a posteriori probability, $G^*(\mathbf{x}) = k^*$, where

$$\begin{aligned} k^* &= \arg\max_k \, p_k(\mathbf{x}) \\ &= \arg\max_k f_{X|Y}(\mathbf{x} \mid k)\pi_k \\ &= \arg\max_k \ln f_{X|Y}(\mathbf{x} \mid k) + \ln \pi_k \end{aligned}$$

or in other words

$$\begin{aligned} p_{k^*}(\mathbf{x}) &\geq p_j(\mathbf{x}) \\ \Leftrightarrow \pi_{k^*} f_{X|Y}(\mathbf{x}|k^*) &\geq \pi_j f_{X|Y}(\mathbf{x} \mid j) \end{aligned} \qquad \text{for all } k^* \neq j.$$

Theorem 2.1 The Bayes rule $G^*(\cdot)$ minimize the error rate (risk),

i.e., $\mathrm{r}(G) \geq \mathrm{r}(G^*)$, for any classifier $G(\cdot)$.

Bayes classifier: binary case

- In the binary problem ($K = 2$), the Bayes classifier reduces to

$$G^*(\mathbf{x}) = \begin{cases} \text{Class 1}, & \text{if } p_2(\mathbf{x}) - p_1(\mathbf{x}) < 0 \\ \text{Class 2}, & \text{if } p_2(\mathbf{x}) - p_1(\mathbf{x}) > 0 \end{cases}$$
$$= \begin{cases} \text{Class 1}, & \text{if } p_2(\mathbf{x}) - \frac{1}{2} < 0 \\ \text{Class 2}, & \text{if } p_2(\mathbf{x}) - \frac{1}{2} > 0. \end{cases}$$

- The case of ties, so when an observation falls in the decision boundary

$$f^*(\mathbf{x}) = p_2(\mathbf{x}) - p_1(\mathbf{x}) = p_2(\mathbf{x}) - 1/2 = 0$$

can be handled by a coin flip.

- Encoding the classes as $\mathcal{G} = \{-1, 1\}$ allows to write the Bayes rule as

$$G^*(\mathbf{x}) = \text{sign}[f^*(\mathbf{x})] \quad \text{with} \quad f^*(\mathbf{x}) = p_1(\mathbf{x}) - \frac{1}{2}.$$

Bayes classifier: binary case

- In the binary problem ($K = 2$), the Bayes classifier reduces to

$$G^*(\mathbf{x}) = \begin{cases} \text{Class 1}, & \text{if } p_2(\mathbf{x}) - p_1(\mathbf{x}) < 0 \\ \text{Class 2}, & \text{if } p_2(\mathbf{x}) - p_1(\mathbf{x}) > 0 \end{cases}$$
$$= \begin{cases} \text{Class 1}, & \text{if } p_2(\mathbf{x}) - \frac{1}{2} < 0 \\ \text{Class 2}, & \text{if } p_2(\mathbf{x}) - \frac{1}{2} > 0. \end{cases}$$

- The case of ties, so when an observation falls in the decision boundary

$$f^*(\mathbf{x}) = p_2(\mathbf{x}) - p_1(\mathbf{x}) = p_2(\mathbf{x}) - 1/2 = 0$$

can be handled by a coin flip.

- Encoding the classes as $\mathcal{G} = \{-1, 1\}$ allows to write the Bayes rule as

$$G^*(\mathbf{x}) = \text{sign}[f^*(\mathbf{x})] \quad \text{with} \quad f^*(\mathbf{x}) = p_1(\mathbf{x}) - \frac{1}{2}.$$

Bayes classifier: binary case

- In the binary problem ($K = 2$), the Bayes classifier reduces to

$$G^*(\mathbf{x}) = \begin{cases} \text{Class 1}, & \text{if } p_2(\mathbf{x}) - p_1(\mathbf{x}) < 0 \\ \text{Class 2}, & \text{if } p_2(\mathbf{x}) - p_1(\mathbf{x}) > 0 \end{cases}$$
$$= \begin{cases} \text{Class 1}, & \text{if } p_2(\mathbf{x}) - \frac{1}{2} < 0 \\ \text{Class 2}, & \text{if } p_2(\mathbf{x}) - \frac{1}{2} > 0. \end{cases}$$

- The case of ties, so when an observation falls in the decision boundary

$$f^*(\mathbf{x}) = p_2(\mathbf{x}) - p_1(\mathbf{x}) = p_2(\mathbf{x}) - 1/2 = 0$$

  can be handled by a coin flip.
- Encoding the classes as $\mathcal{G} = \{-1, 1\}$ allows to write the Bayes rule as

$$G^*(\mathbf{x}) = \text{sign}[f^*(\mathbf{x})] \quad \text{with} \quad f^*(\mathbf{x}) = p_1(\mathbf{x}) - \frac{1}{2}.$$

Bayes classifier: binary case (cont'd)

- The decision regions can be expressed as

$$\Gamma_0^* = \left\{ \mathbf{x} : \frac{f_{X|Y}(\mathbf{x} \mid 0)}{f_{X|Y}(\mathbf{x} \mid 1)} > \frac{\pi_1}{\pi_0} \right\}$$
$$\Gamma_1^* = \left\{ \mathbf{x} : \frac{f_{X|Y}(\mathbf{x} \mid 0)}{f_{X|Y}(\mathbf{x} \mid 1)} < \frac{\pi_1}{\pi_0} \right\}$$

  and handling ties as random guessing.

- The detection rule can be expressed as

$$L(\mathbf{x}) = \frac{f_{X|Y}(\mathbf{x} \mid 0)}{f_{X|Y}(\mathbf{x} \mid 1)} \underset{1}{\overset{0}{\gtrless}} \frac{\pi_1}{\pi_0}$$

  and notice that $L(\mathbf{x})$ is the likelihood ratio.

Bayes classifier: binary case (cont'd)

- The decision regions can be expressed as

$$\Gamma_0^* = \left\{ \mathbf{x} : \frac{f_{X|Y}(\mathbf{x} \mid 0)}{f_{X|Y}(\mathbf{x} \mid 1)} > \frac{\pi_1}{\pi_0} \right\}$$
$$\Gamma_1^* = \left\{ \mathbf{x} : \frac{f_{X|Y}(\mathbf{x} \mid 0)}{f_{X|Y}(\mathbf{x} \mid 1)} < \frac{\pi_1}{\pi_0} \right\}$$

  and handling ties as random guessing.
- The detection rule can be expressed as

$$L(\mathbf{x}) = \frac{f_{X|Y}(\mathbf{x} \mid 0)}{f_{X|Y}(\mathbf{x} \mid 1)} \underset{1}{\overset{0}{\gtrless}} \frac{\pi_1}{\pi_0}$$

  and notice that $L(\mathbf{x})$ is the likelihood ratio.

Esimerkki 2.1

Assume $K = 2$ two classes with class conditional distributions following exponential distributions:

$$X|Y = 0 \sim \mathrm{Exp}(\lambda_0) \quad \text{and} \quad X|Y = 1 \sim \mathrm{Exp}(\lambda_1).$$

Hence

$$f_{X|Y}(x|k) = \lambda_k \exp\{-\lambda_k x\},\ x > 0$$

where $\lambda_k > 0$ is the rate parameter, $\lambda_k^{-1} = \mathbb{E}[X \mid Y = k]$, $k \in \{0, 1\}$. W.l.o.g. assume $\lambda_0 < \lambda_1$.

a) Derive the classification region Γ_0^* (that minimize the Bayes risk) when $\pi_0 = \pi_1 = 1/2$.

b) Calculate the risk (probability of an error) $\mathrm{r}(G^*) = \Pr(Y \neq G^*(X))$ when $\lambda_0 = 1$ and $\lambda_1 = 3$.

Menu

Classification costs

- Cost function quantifies the consequences of the decisions

$$C(k, j) : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$$

so assigns a cost of misclassifying an observation into class j when its true class is k.

- Commonly, one assumes

$$C(k, k) = 0 \qquad \text{and} \qquad C(k, j) > 0, \quad \forall k \neq j \in \mathcal{G}$$

i.e., 0 cost for correct classification and non-zero otherwise.

- *Uniform cost*

$$C(k, j) = 1_{\{k \neq j\}} = \begin{cases} 1, & k \neq j \\ 0, & k = j \end{cases}$$

gives unit cost to all errors.

Classification costs

- Cost function quantifies the consequences of the decisions

$$C(k, j) : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$$

so assigns a cost of misclassifying an observation into class j when its true class is k.

- Commonly, one assumes

$$C(k, k) = 0 \qquad \text{and} \qquad C(k, j) > 0, \quad \forall k \neq j \in \mathcal{G}$$

i.e., 0 cost for correct classification and non-zero otherwise.

- *Uniform cost*

$$C(k, j) = 1_{\{k \neq j\}} = \begin{cases} 1, & k \neq j \\ 0, & k = j \end{cases}$$

gives unit cost to all errors.

Classification costs

- Cost function quantifies the consequences of the decisions

$$C(k, j) : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$$

so assigns a cost of misclassifying an observation into class j when its true class is k.

- Commonly, one assumes

$$C(k, k) = 0 \qquad \text{and} \qquad C(k, j) > 0, \quad \forall k \neq j \in \mathcal{G}$$

i.e., 0 cost for correct classification and non-zero otherwise.

- *Uniform cost*

$$C(k, j) = 1_{\{k \neq j\}} = \begin{cases} 1, & k \neq j \\ 0, & k = j \end{cases}$$

gives unit cost to all errors.

Expected cost of misclassification

- The expected cost of misclassification of rule $G(\cdot)$ is

$$\begin{aligned}\mathsf{ECM}(G) &= \mathbb{E}_{X,Y}\big[C\big(Y, G(X)\big)\big] \\ &= \sum_{k=1}^{K} \mathbb{E}_{X|Y}\big[C\big(k, G(X)\big)\big|Y = k\big]\pi_k,\end{aligned}$$

- ECM is equal to Bayes risk when uniform cost is used.
- In the binary classification problem ($\mathcal{G} = \{0, 1\}$), the ECM is

$$\pi_0 \cdot C(0,1) \int_{\Gamma_1(G)} f_{X|Y}(\mathbf{x} \mid 0)\mathrm{d}\mathbf{x} + \pi_1 \cdot C(1,0) \int_{\Gamma_0(G)} f_{X|Y}(\mathbf{x} \mid 1)\mathrm{d}\mathbf{x}.$$

Expected cost of misclassification

- The expected cost of misclassification of rule $G(\cdot)$ is

$$\begin{aligned} \mathsf{ECM}(G) &= \mathbb{E}_{X,Y}\big[C\big(Y, G(X)\big)\big] \\ &= \sum_{k=1}^{K} \mathbb{E}_{X|Y}\big[C\big(k, G(X)\big)\big|Y = k\big]\pi_k, \end{aligned}$$

- ECM is equal to Bayes risk when uniform cost is used.
- In the binary classification problem ($\mathcal{G} = \{0, 1\}$), the ECM is

$$\pi_0 \cdot C(0,1) \int_{\Gamma_1(G)} f_{X|Y}(\mathbf{x} \mid 0)\mathrm{d}\mathbf{x} + \pi_1 \cdot C(1,0) \int_{\Gamma_0(G)} f_{X|Y}(\mathbf{x} \mid 1)\mathrm{d}\mathbf{x}.$$

Expected cost of misclassification

- The expected cost of misclassification of rule $G(\cdot)$ is

$$\begin{aligned}\mathsf{ECM}(G) &= \mathbb{E}_{X,Y}\big[C\big(Y, G(X)\big)\big] \\ &= \sum_{k=1}^{K} \mathbb{E}_{X|Y}\big[C\big(k, G(X)\big)\big|Y = k\big]\pi_k,\end{aligned}$$

- ECM is equal to Bayes risk when uniform cost is used.
- In the binary classification problem ($\mathcal{G} = \{0, 1\}$), the ECM is

$$\pi_0 \cdot C(0,1) \int_{\Gamma_1(G)} f_{X|Y}(\mathbf{x} \mid 0)\mathrm{d}\mathbf{x} + \pi_1 \cdot C(1,0) \int_{\Gamma_0(G)} f_{X|Y}(\mathbf{x} \mid 1)\mathrm{d}\mathbf{x}.$$

Minimizing the ECM

Theorem. The discriminant rule $G^*(\mathbf{x})$ that minimizes the ECM is based on discriminant scores

$$G_1(\mathbf{x}) = \ln f_{X|Y}(\mathbf{x}|1) + \ln \pi_1 + \ln C(1,2),$$
$$G_2(\mathbf{x}) = \ln f_{X|Y}(\mathbf{x}|2) + \ln \pi_2 + \ln C(2,1),$$

where the decision regions are

$$\Gamma_1^* = \{\mathbf{x} : G_1(\mathbf{x}) \geq G_2(\mathbf{x})\} = \left\{\mathbf{x} : \frac{f_{X|Y}(\mathbf{x}|1)}{f_{X|Y}(\mathbf{x}|2)} \geq \frac{C(2,1)}{C(1,2)} \cdot \frac{\pi_2}{\pi_1}\right\},$$
$$\Gamma_2^* = \{\mathbf{x} : G_1(\mathbf{x}) < G_2(\mathbf{x})\} = \left\{\mathbf{x} : \frac{f_{X|Y}(\mathbf{x}|1)}{f_{X|Y}(\mathbf{x}|2)} < \frac{C(2,1)}{C(1,2)} \cdot \frac{\pi_2}{\pi_1}\right\}.$$

- Choosing uniform costs, one obtain the Bayes rule.

Example 2.2

- Suppose it is known that it is twice as costly to assign an observation from class 1 to class 0 than vice versa and that approximately 20% of observations belong to class 1.
- When an observation $\mathbf{x}$ receives values $f_{X|Y}(\mathbf{x} \mid 0) = 0.3$ and $f_{X|Y}(\mathbf{x} \mid 1) = 0.4$, then is it classified to class 1 or class 2?

Menu

Margin

- If we encode $Y \in \mathcal{G} = \{-1, 1\}$, then a classifier G can be expressed as

$$G(\mathbf{x}) = \mathsf{sign}[f(\mathbf{x})]$$

  where $f : \mathbb{R}^p \to \mathbb{R}$ is called the predictor function.
- The decision boundary is defined by $f(\mathbf{x}) = 0$.
- We can express the misclassification loss as

$$L_{0/1}(y, f(\mathbf{x})) = 1_{\{y \neq G(\mathbf{x})\}} = 1_{\{y\mathsf{sign}[f(\mathbf{x})] \neq 1\}} = 1_{\{yf(\mathbf{x})<0\}}$$

  where $m = yf(\mathbf{x})$ is called the margin of $(y, \mathbf{x})$.
- The risk can be expressed as $\mathrm{r}(f) = \mathbb{E}\big[1_{\{Yf(X)<0\}}\big]$.
- Margin $m = yf(\mathbf{x})$ is useful since
  - $y_i f(\mathbf{x}_i) > 0 \Rightarrow \mathbf{x}_i$ is classified correctly
  - $y_i f(\mathbf{x}_i) < 0 \Rightarrow \mathbf{x}_i$ is misclassified.

Loss functions

- If would be natural to find predictor function f that minimizes the associated empirical risk $\frac{1}{N}\sum_{i=1}^{N} 1_{\{y_i f(\mathbf{x}_i)<0\}}$.

  ... but this problem turns out to be NP-complete
- Thus we find a function $f(\mathbf{x})$ that optimizes a loss function

$$L(y, f(\mathbf{x})) : \mathcal{G} \times \mathbb{R} \to \mathbb{R}$$

  using some other loss function than $1/0$-loss.
- Commonly $L(\cdot,\cdot)$ will be a function of margin $yf(\mathbf{x})$ only.
- We choose loss fnc $L(\cdot,\cdot)$ and determine $f^*(\mathbf{x})$ associated with it

$$f^*(\mathbf{x}) = \arg\min_{f(\mathbf{x})} \mathbb{E}\big[L\big(Y, f(\mathbf{x})\big) \mid X = \mathbf{x}\big]$$

  where f^* should produce positive margins as frequently as possible.
- $G(\mathbf{x}) = \text{sign}[f^*(\mathbf{x})]$ is then the classification rule.

Loss functions (cont'd)

$$f^*(\mathbf{x}) = \arg\min_{f(\mathbf{x})} \mathbb{E}\big[L\big(Y, f(\mathbf{x})\big) \mid X = \mathbf{x}\big]$$

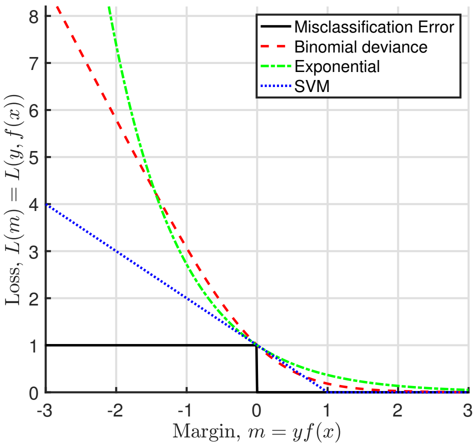
name	$L(y, f(\mathbf{x}))$	$f^*(\mathbf{x})$
misclassification loss	$1_{\{yf(\mathbf{x})<0\}}$	$p(\mathbf{x}) - 1/2$
exponential loss	$\exp(-yf(\mathbf{x}))$	$\frac{1}{2}\log\frac{p(\mathbf{x})}{1-p(\mathbf{x})}$
binomial deviance	$\log\left(1 + e^{-2y\,f(\mathbf{x})}\right)$	$\frac{1}{2}\log\frac{p(\mathbf{x})}{1-p(\mathbf{x})}$
support vector (hinge loss)	$[1 - yf(\mathbf{x})]_+$	$p(\mathbf{x}) - 1/2$

Table. Popular loss functions for classification as well as the associated optimal predictor function. We use shorthand notation $p(\mathbf{x})$ for $p_1(\mathbf{x}) = \Pr(Y = 1 \mid X = \mathbf{x})$.

Notation $[y]_+$ means positive part of y, i.e., $[y]_+ = y1_{\{y>0\}} = \max(y, 0)$.

Loss functions for classification

Loss functions when displayed as a function of margin $m = yf(\mathbf{x})$:



- **Misclassification**: $1_{\{m<0\}}$;
- **Exponential loss**: $\exp(-m)$
- **Binomial deviance**: $\log(1 + \exp(-2m))$
- **Support vector**: $(1 - m)_+$

Loss function comparisons

- All loss fnc-s above are monotone decreasing continuous approximations (or upperbounds) to misclassification loss.
- Misclassification loss is non-differentiable and non-convex in margin m and not suitable for optimization.
- The difference between loss functions is in degree of penalizing negative margins.
- Binomial deviance is more robust than exponential loss:
  - it assigns less influence on observations with large negative margins.
  - it grows linearly as the margin value m tends to $-\infty$.

  This yields more robustness to misslabelling also.

Binomial deviance loss function

- The conditional probability mass fnc of $Y'|X = \mathbf{x} \sim \mathrm{Ber}(p(\mathbf{x}))$ is

$$f_{Y'|X}(y'|\mathbf{x}) = p(\mathbf{x})^{y'}(1-p(\mathbf{x}))^{1-y'}, \quad y' \in \{0,1\}.$$

- Its log-likelihood function is

$$l(y', p(\mathbf{x})) = y' \log p(\mathbf{x}) + (1-y')\log(1-p(\mathbf{x})), \quad y' \in \{0,1\}.$$

  which is also sometimes referred to as cross entropy loss.

- Binomial deviance is its negative log-likelihood which may be written as

$$-l(y, f(\mathbf{x})) = \log(1 + e^{-2yf(\mathbf{x})}), \quad y \in \{-1,1\}.$$

  using output encoding $y = 2y' - 1 \in \{-1,1\}$ and symmetric logistic transformation:

$$f(\mathbf{x}) = \frac{1}{2}\log\frac{p(\mathbf{x})}{1-p(\mathbf{x})} \Leftrightarrow p(\mathbf{x}) = \frac{e^{f(\mathbf{x})}}{e^{f(\mathbf{x})} + e^{-f(\mathbf{x})}} = \frac{1}{1+e^{-2f(\mathbf{x})}}$$

Menu

Performance measures

- Confusion matrix:

		Predicted class $+$	$-$	sum
True class	$+$	TP	FN	TP $+$ FN
	$-$	FP	TN	FP $+$ TN

FN (false negatives) aka miss-detection

$=$ # of observations of class $+$ that are misclassified

FP (false positives) aka false alarm

$=$ # of observations of class $-$ that are misclassified

- Recall aka true positive rate

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{\text{\#found true objects}}{\text{\#all true objects}}$$

Performance measures (cont´d)

- Specificity aka true negative rate (TNR):

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{FP} + \mathrm{TN}}.$$

- Precision (emphasizes TP-s and FP-s)

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = \frac{\#\text{found true objects}}{\#\text{found all objects}}$$

- F1-score is harmonic mean of precision and recall metrics:

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

Test error rate

1. Hold-out: Randomly split the available data into training set and test set using some split ratio.
2. Training set is used to construct the classifier $\hat{G}$ and the test set is used to evaluate a performance measure on a test set.
3. Repeat steps 1 and 2 (using another random split), e.g., 100 times.
4. TER is computed as the average of obtained error rates while the test set accuracy is computed as $1 - \mathrm{TER}$.

Menu

Conclusions

- Basic concepts of classification were introduced
  - Bayes risk and 0/1-loss
  - margin, predictor function and loss functions (including exponential loss, binomial deviance, etc)
  - cost of misclassification and the expected cost of misclassification (ECM)
- How to compare classification performance using different performance measures
  - confusion matrix
  - precision, recall, F1-score, specificity, etc.
  - test error rate (TER)
- Although the concepts were illustrated mostly in binary classification problem, they generalize to multi-class case.