Deep Learning (BEV033DLE)
Lecture 1.

Czech Technical University in Prague

- Predictors, Risk, Empirical risk
- Learning predictors
- Generalisation bounds

Organisational Matters

Teachers: Alexander Shekhovtsov, Boris Flach

Format: 1 lecture & 1 lab per week (6 credits), labs of two types (alternating)

- practical labs: implementation of selected methods (Python)
- theoretical labs: solving theoretical assignments

Grading: 40% practical labs + 60% written exam = 100% (+ bonus points)

Prerequisites:

- calculus, linear algebra and optimisation
- basics of graph theory and related algorithms
- pattern recognition and machine learning (AE4B33RPZ)

More details: `https://cw.fel.cvut.cz/b192/courses/bev033dle/start`

Predictors, Risk, Empirical risk

- **object features** $x \in \mathcal{X}$, can be categorical, numerical, vectors, etc.
- **state of the object** $y \in \mathcal{Y}$ is usually hidden
- **prediction strategy** $h\colon \mathcal{X} \to \mathcal{Y}$ predicts the hidden state $y = h(x)$ given the features x.

Example 1. Consider the following predictors

a) $x =$  $\xRightarrow{h}$ $y =$ apple $\quad \mathcal{Y} = \{\text{apple, pear}, \ldots\}$ - classes

b) $x =$  $\xRightarrow{h}$ $y = 45$ $\quad \mathcal{Y} = \{0, 1, \ldots, 150\} \subset \mathbb{N}$ - age

c) $x =$ 

$\xRightarrow{h}$ $y = (y_1, \ldots, y_T)$ $\quad y \in \mathbb{Z}^{2T}$ - trajectory

> Q: How to measure the quality of a predictor?

Predictors, Risk, Empirical risk

- **loss function** $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ penalises wrong predictions,
  i.e. $\ell(y, h(x))$ is the loss for predicting $y' = h(x)$ when y is the true state

Example 1. cont'd

a) $\ell(y, y') = \mathbb{1}\{y \neq y'\}$, i.e. simple 0/1 loss

b) $\ell(y, y') = (y - y')^2$, i.e. squared error

c) $\ell(y, y') = \sum_t \|y_t - y'_t\|^2$, i.e. squared trajectory distance

Main assumption: x and y are random variables, related by a joint but *unknown* probability distribution $p(x, y)$.

Measuring the quality of a predictor: draw independent pairs $x, y \sim p(x, y)$ infinitely often and compute the expected loss $\Rightarrow$ Risk of the predictor

$$R(h) = \mathbb{E}_{x,y \sim p(x,y)}\big[\ell\big(y, h(x)\big)\big] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y)\,\ell\big(y, h(x)\big)$$

But we don't know $p(x, y)$ and don't have infinite time!

Predictors, Risk, Empirical risk

Practical approach: Empirical risk for an i.i.d. test sample $\mathcal{T}^m = \{(x^j, y^j) \mid j = 1, \dots, m\}$

$$R_{\mathcal{T}^m}(h) = \frac{1}{m} \sum_{j=1}^{m} \ell\big(y^j, h(x^j)\big)$$

Generalisation error: How strong can $R_{\mathcal{T}^m}(h)$ deviate from $R(h)$?

$$\mathcal{T}^m \sim p(x, y) \;\Rightarrow\; \mathbb{P}\Big(|R(h) - R_{\mathcal{T}^m}(h)| > \varepsilon\Big) < ??$$

- Chebyshev inequality $\Rightarrow \mathbb{P}\Big(|R(h) - R_{\mathcal{T}^m}(h)| > \varepsilon\Big) < \frac{\mathbb{V}[\ell(y, h(x))]}{m\varepsilon^2}$,
  loose bound, requires to know $\mathbb{V}[\ell(y, h(x))]$, converges slowly for $m \to \infty$.
- Hoeffding inequality $\Rightarrow \mathbb{P}\Big(|R(h) - R_{\mathcal{T}^m}(h)| > \varepsilon\Big) < 2e^{-\frac{2m\varepsilon^2}{(\triangle\ell)^2}}$,
  where $\triangle\ell = \ell_{max} - \ell_{min}$.

Example 2. Consider a classifier with $0/1$ loss. What test set size m ensures that $R_{\mathcal{T}^m}(h) - 0.01 < R(h) < R_{\mathcal{T}^m}(h) + 0.01$ with probability 95%?

Answer: $m \approx 2 \cdot 10^4$.

Learning predictors

Generative learning: Specify a class of distributions $p_\theta(x,y)$, $\theta \in \Theta$, collect an i.i.d. training set $\mathcal{T}^m$, estimate θ_* e.g. by maximal likelihood estimator and then predict by

$$h(x) = \operatorname*{arg\,min}_{y \in \mathcal{Y}} \sum_{y' \in \mathcal{Y}} p_{\theta_*}(x, y') \ell(y', y)$$

Discriminative learning: Specify a hypothesis class $\mathcal{H}$ of predictors, collect an i.i.d. training set $\mathcal{T}^m$, select the predictor $h_m \in \mathcal{H}$ that minimises the empirical risk

$$h_m = \operatorname*{arg\,min}_{h \in \mathcal{H}} \frac{1}{m} \sum_{j=1}^{m} \ell\big(y^j, h(x^j)\big)$$

Question: Can we bound the estimation error $R(h_m) - R(h_{\mathcal{H}})$, where

$$h_{\mathcal{H}} = \operatorname*{arg\,min}_{h \in \mathcal{H}} R(h)$$

denotes the best predictor from $\mathcal{H}$?

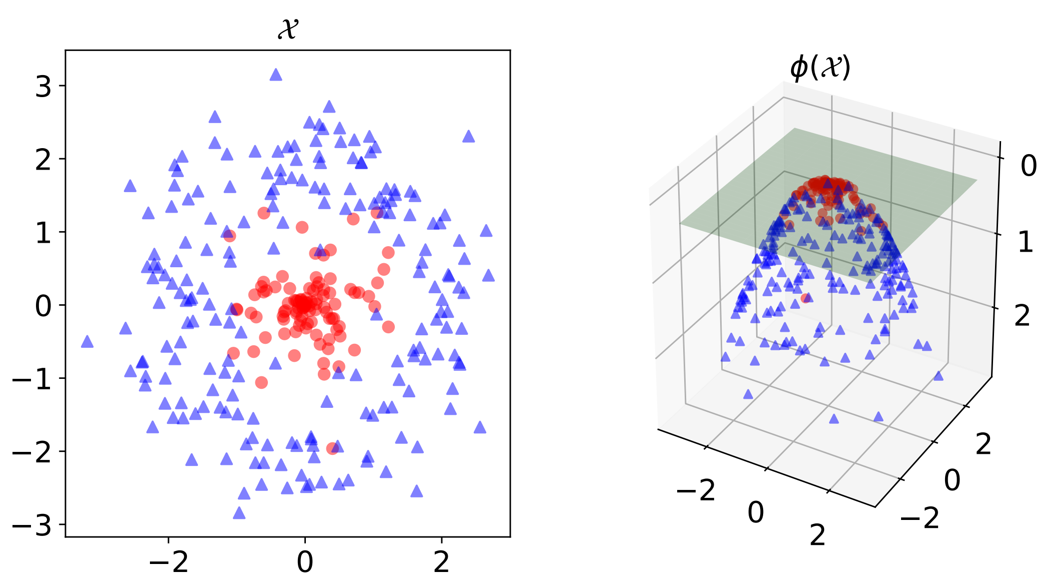
Learning predictors

Example 3. (Linear classifier)
Consider a binary classifier, i.e. $|\mathcal{Y}| = 2$ and 0/1 loss

- Encode the two classes by $y = \pm 1$
- Define a mapping $\phi\colon \mathcal{X} \to \mathbb{R}^n$. If $\mathcal{X} = \mathbb{R}^n$, this mapping can be the identity mapping I.
- Define the hypothesis class by $\mathcal{H}$

$$y = \operatorname{sign}\big[\langle w, \phi(x)\rangle + b\big]$$

where $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are parameters.
We can get rid of b by defining $\phi'\colon \mathcal{X} \to \mathbb{R}^{n+1}$ by $\phi'(x) = (\phi(x), 1)$.



Learning predictors

Example 3. (cont'd)
Given i.i.d. training data $\mathcal{T}^m = \big\{(x^j, y^j) \,\big|\, j = 1, \ldots, m\big\}$, we want to learn the classifier by empirical risk minimisation. This amounts to solve

$$R_{\mathcal{T}^m}(h_w) = \frac{1}{m}\sum_{j=1}^{m} \ell\big(y^j, \operatorname{sign}\big\langle w, \phi(x^j)\big\rangle\big) = \frac{1}{m}\sum_{j=1}^{m} \mathrm{H}\big(-y^j\big\langle w, \phi(x^j)\big\rangle\big) \to \min_{w},$$

where H denotes the Heaviside function. **Objection:** But this task is not tractable!

Ways out:

- Redefine the loss in terms of $\gamma = y\langle w, \phi(x)\rangle$, i.e. $\mathrm{H}(-\gamma)$ and replace it by *hinge loss*. Combined with L_2 regularisation, this leads to SVMs and a convex optimisation task.
- A second approach has an interpretation in terms of a statistical model known as "logistic regression". We assume

$$p_w(y \mid x) = \frac{e^{y\langle w, \phi(x)\rangle}}{e^{\langle w, \phi(x)\rangle} + e^{-\langle w, \phi(x)\rangle}} = \frac{1}{1 + e^{-2y\langle w, \phi(x)\rangle}}$$

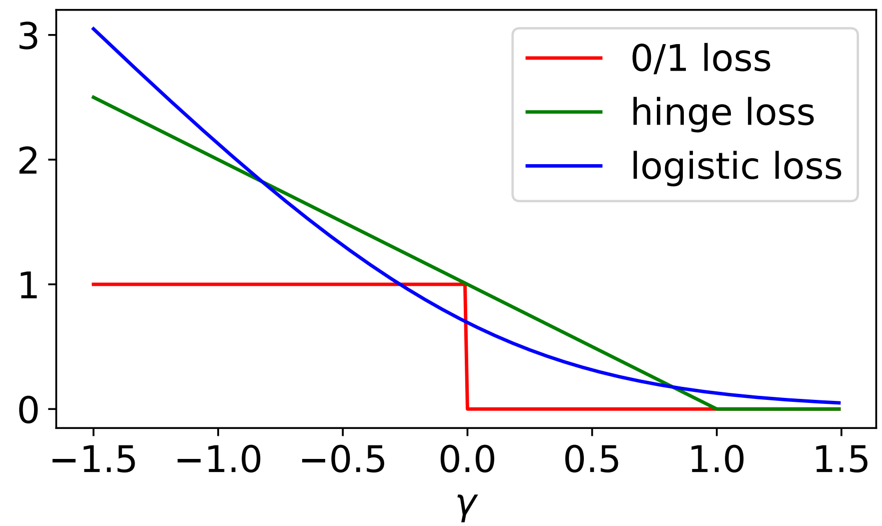
Learning predictors

Example 3. (cont'd)

and maximise the expected conditional log-likelihood of the training data $\mathcal{T}^m$

$$\frac{1}{m}\sum_{j=1}^{m}\log p_w(y^j \mid x^j) = -\frac{1}{m}\sum_{j=1}^{m}\log\Big[1+e^{-2y^j\langle w,\phi(x^j)\rangle}\Big] \to \max_w$$

This is a concave optimisation task. Notice, we do not model a joint distribution $p(x,y)$, but conditional distributions $p(y \mid x)$ only.



Generalisation bounds

Generalisation bounds: If we learn a predictor $h_m \in \mathcal{H}$ by (surrogate) empirical risk minimisation on training data $\mathcal{T}^m = \big\{(x^j, y^j) \,\big|\, j = 1, \ldots, m\big\}$, how probable is it that the obtained predictor will be close to the optimal one? I.e.

$$\mathbb{P}\Big(|R(h_{\mathcal{H}}) - R(h_m)| > \varepsilon\Big) < ??$$

For **binary classifiers**, i.e. $|\mathcal{Y}| = 2$ and 0/1-loss, this question is answered as follows.

Definition 1. Let $M = \{x^j \in \mathcal{X} \mid j = 1, \ldots, m\}$ be a set of input observations and $\mathcal{H}$ be a set of binary classifiers. The set M is said to be *shattered* by $\mathcal{H}$, if there exists a predictor $h \in \mathcal{H}$ for each possible classification $\boldsymbol{y} \in \{-1, +1\}^m$ of M, s.t. $y^j = h(x^j)$, $\forall j = 1 \ldots, m$. The Vapnik-Chervonenkis dimension of $\mathcal{H}$ is the cardinality of the largest subset of $\mathcal{X}$ shattered by $\mathcal{H}$.

Theorem 1. *Let $\mathcal{H}$ be a set of binary classifiers with VC-dimension d and $\mathcal{T}^m$ be an i.i.d training set drawn from $p(x, y)$. Then*

$$\mathbb{P}\Big(\sup_{h \in \mathcal{H}} |R(h) - R_{\mathcal{T}^m}(h)| > \varepsilon\Big) < 4\Big(\frac{2em}{d}\Big)^d e^{\frac{-m\varepsilon^2}{8}}$$

holds for any $\varepsilon > 0$.

Generalisation bounds

It follows that empirical risk minimisation is statistically consistent for binary valued predictor sets $\mathcal{H}$ with finite VC-dimensions.

Corollary 1. *A set $\mathcal{H}$ of binary predictors with finite VC-dimension satisfies*

$$\lim_{m\to\infty} \mathbb{P}\Big(\sup_{h\in\mathcal{H}} |R(h) - R_{\mathcal{T}^m}(h)| > \varepsilon\Big) = 0.$$

Corollary 2. *If a set $\mathcal{H}$ of binary predictors has finite VC-dimension, then empirical risk minimisation is consistent in $\mathcal{H}$, i.e.*

$$\lim_{m\to\infty} \mathbb{P}\Big(R(h_m) - R(h_{\mathcal{H}}) \geqslant \varepsilon\Big) = 0$$

for any $\varepsilon > 0$.

All this covers ERM for binary valued predictors only.

What about non-binary classifiers, regression etc?

Take home messages

- predictors, loss function, risk,
- risk vs. empirical risk, Hoeffding inequality,
- two approaches for learning: generative learning and empirical risk minimisation,
- empirical risk minimisation for linear classifiers is hard; ways out: SVMs, logistic regression,
- finite VC dimension of a class of predictors ensures consistency of learning and provides worst case generalisation bounds.