Mathematical techniques in data science

PAC Learning

Supervised learning: standard setting

- Given: a sequence of label data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ sampled (independently and identically) from an unknown distribution $P_{X,Y}$
- a learning algorithm seeks a function $h : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space

Supervised learning: standard setting

- The function *h* is an element of some space of possible functions $\mathcal{H}$, usually called the *hypothesis space*
- In order to measure how well a function fits the training data, a *loss function*

$$L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^{\geq 0}$$

  is defined

Risk and empirical risk

- With a pre-defined loss function, the “optimal hypothesis” is the minimizer over $\mathcal{H}$ of the risk function

$$R(h) = E_{(X,Y)\sim P}[L(Y, h(X))]$$

- Since P is unknown, the simplest approach is to approximate the risk function by the empirical risk

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, h(x_i))$$

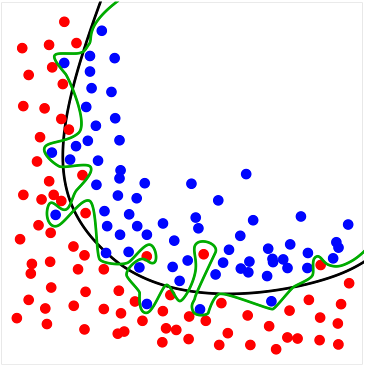
- The empirical risk minimizer (ERM): minimizer of the empirical risk function (in this lecture, denoted by $\hat{h}_n$)
- Let h^* denotes a minimizer of the risk function

Failure of ERM

We hope that

$$R(\hat{h}_n) \approx R(h^*),$$

but in general, this might not be true if the hypothesis space $\mathcal{H}$ is too large



Failure of ERM

- We hope that
$$R(\hat{h}_n) \approx R(h^*),$$
but in general, this might not be true if the hypothesis space $\mathcal{H}$ is too large
- Question: What does "too large" mean?
- We need to be able to quantify/control the difference between $R(\hat{h}_n)$ and $R(h^*)$

Modes of estimations

- Analysis

$$\lim_{n\to\infty} x_n = x$$

- Numerical analysis

$$\|x_n - x\| = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \quad \text{or} \quad \|x_n - x\| \leq \frac{C}{\sqrt{n}}$$

- PAC (Probably Approximately Correct) learning

$$\|x_n - x\| \leq C(\delta)\frac{1}{\sqrt{n}}$$

with probability at least $1 - \delta$

PAC learning

Definition

The probably approximately correct (PAC) learning model typically states as follows: we say that $\hat{h}_n$ is ϵ-accurate with probability $1 - \delta$ if

$$P\left[R(\hat{h}_n) - R(h^*) > \epsilon\right] < \delta.$$

In other words, we have $R(\hat{h}_n) - R(h^*) \leq \epsilon$ with probability at least $(1 - \delta)$.

Probability inequalities

Markov inequality

Theorem (Markov inequality)

For any nonnegative random variable X and $\epsilon > 0$,

$$P[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon}.$$

Markov inequality

Theorem

For any random variable X, $\epsilon > 0$ *and* $t > 0$

$$P[X \geq \epsilon] \leq \frac{\mathbb{E}[e^{tX}]}{e^{t\epsilon}}.$$

Exponential moment of bounded random variables

Theorem

If random variable X has mean zero and is bounded in $[a, b]$, then for any $s > 0$,

$$\mathbb{E}[e^{tX}] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right)$$

Hoeffding's inequality

Theorem (Hoeffding's inequality)

Let $X_1, X_2, \ldots, X_n$ be i.i.d copy of a random variable $X \in [a, b]$, and $\epsilon > 0$,

$$P\left[\frac{X_1 + X_2 + \ldots + X_n}{n} - E[X] \geq \epsilon\right] \leq \exp\left(-\frac{n\epsilon^2}{2(b-a)^2}\right).$$

Corollary:

$$P\left[\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - E[X]\right| \geq \epsilon\right] \leq 2\exp\left(-\frac{n\epsilon^2}{2(b-a)^2}\right).$$

Generalization bound for finite hypothesis space and bounded loss

Assumption

- the loss function L is bounded, that is

$$0 \leq L(y, y') \leq c \quad \forall y, y' \in \mathcal{Y}$$

- the hypothesis space is a finite set, that is

$$\mathcal{H} = \{h_1, h_2, \ldots, h_m\}.$$

Key ideas

- For any $h \in \mathcal{H}$ and $\epsilon > 0$ we have

$$P[|R_n(h) - R(h)| \geq \epsilon] \leq 2\exp\left(-\frac{n\epsilon^2}{2c^2}\right).$$

- Using a union bound on the "failure probability" associated with each hypothesis, we have

$$P[\exists h \in \mathcal{H} : |R_n(h) - R(h)| \geq \epsilon] \leq 2|\mathcal{H}|\exp\left(-\frac{n\epsilon^2}{2c^2}\right).$$

Key ideas

- Using a union bound on the "failure probability" associated with each hypothesis, we have

$$P[\forall h \in \mathcal{H} : |R_n(h) - R(h)| < \epsilon]$$
$$\geq 1 - 2|\mathcal{H}| \exp\left(-\frac{n\epsilon^2}{2c^2}\right).$$

- Under this "good event":

$$\begin{aligned} & R(\hat{h}_n) - R(h^*) \\ = & [R(\hat{h}_n) - R_n(\hat{h}_n)] + [R_n(\hat{h}_n) - R_n(h^*)] + [R_n(h^*) - R(h^*)] \\ \leq & 2\epsilon \end{aligned}$$

- Conclusion: $\hat{h}_n$ is (2ϵ)-accurate with probability $1 - \delta$, where

$$\delta = 2|\mathcal{H}| \exp\left(-\frac{n\epsilon^2}{2c^2}\right)$$

PAC estimate for ERM

Theorem

For any $\delta > 0$ *and* $\epsilon > 0$*, if*

$$n \geq \frac{8c^2}{\epsilon^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right)$$

then $\hat{h}_n$ *is* ϵ*-accurate with probability at least* $1 - \delta$*.*

PAC estimate for ERM

$$n = \frac{8c^2}{\epsilon^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right)$$

- Fix a level of confidence δ, the accuracy ϵ of the ERM is

$$\mathcal{O}\left(\frac{1}{\sqrt{n}}\sqrt{\log\left(\frac{1}{\delta}\right) + \log(|\mathcal{H}|)}\right)$$

- If we want $\epsilon \to 0$ as $n \to \infty$:

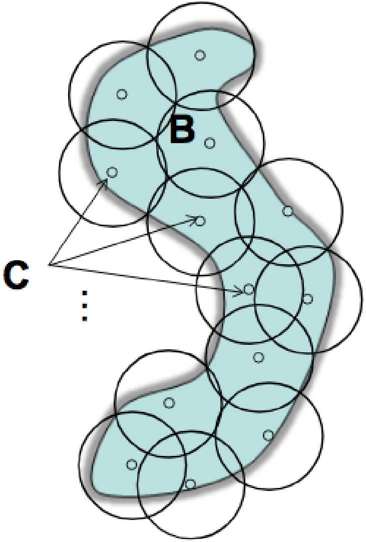
$$\log(|\mathcal{H}|) \ll n$$

- The convergence rate will not be better than $\mathcal{O}(n^{-1/2})$

Generalization bound using covering number.

Covering numbers

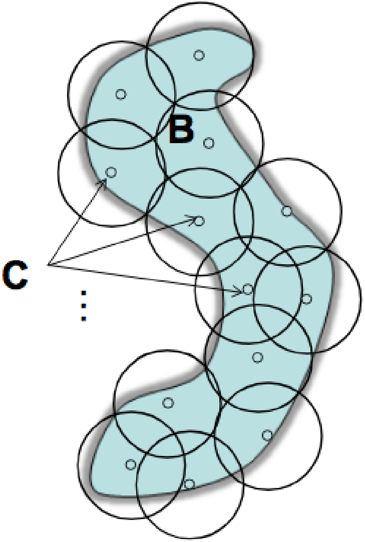
- Assumption: $\mathcal{H}$ is a metric space with distance d defined on it.
- For $\epsilon > 0$, we denote by $\mathcal{N}(\epsilon, \mathcal{H}, d)$ the *covering number* of $(\mathcal{H}, d)$; that is, $\mathcal{N}(\epsilon, \mathcal{H}, d)$ is the minimal number of balls of radius ϵ needed to cover $\mathcal{H}$.



Covering numbers

Remark: If $\mathcal{H}$ is a bounded $k-$dimensional manifold/algebraic surface, then we now that

$$\mathcal{N}(\epsilon, \mathcal{H}, d) = \mathcal{O}\left(\epsilon^{-k}\right)$$



Generalization bound using covering number.

- Assumption: $\mathcal{H}$ is a metric space with distance d defined on it.
- For $\epsilon > 0$, we denote by $\mathcal{N}(\epsilon, \mathcal{H}, d)$ the *covering number* of $(\mathcal{H}, d)$; that is, $\mathcal{N}(\epsilon, \mathcal{H}, d)$ is the minimal number of balls of radius ϵ needed to cover $\mathcal{H}$.
- Assumption: loss function L satisfies:

$$|L(h(x), y) - L(h'(x), y)| \leq Cd(h, h') \quad \forall, x \in \mathcal{X}; y \in \mathcal{Y}; h, h' \in \mathcal{H}$$

Key ideas

- If
$$n = \frac{8c^2}{\epsilon^2} \log\left(\frac{2|\mathcal{H}_\epsilon|}{\delta}\right)$$
then the event
$$|R_n(h) - R(h)| \leq \epsilon, \forall h \in \mathcal{H}_\epsilon$$
happens with probability at least $1 - \delta$.
- Under this event, consider any $h \in \mathcal{H}$, then there exists $h_0 \in \mathcal{H}_\epsilon$ such that $d(h, h_0) \leq \epsilon$.

Key ideas

- Since the loss function is Lipschitz

$$|R_n(h) - R_n(h_0)| \leq Cd(h, h_0)$$

  and

$$|R(h) - R(h_0)| \leq Cd(h, h_0).$$

- Conclusion:

$$|R_n(h) - R(h)| \leq (2C + 1)\epsilon \quad \forall h \in \mathcal{H}.$$

Generalization bound using covering number.

Theorem

For all $\epsilon > 0$, $\delta > 0$, *if*

$$n \geq \frac{c^2}{2\epsilon^2} \log\left(\frac{2\mathcal{N}(\epsilon, \mathcal{H}, d)}{\delta}\right)$$

then

$$|R_n(h) - R(h)| \leq (2C + 1)\epsilon \quad \forall h \in \mathcal{H}.$$

with probability at least $1 - \delta$.

Example: Polynomial covering number.

- Assume that

$$\mathcal{N}(\epsilon, \mathcal{H}, d) \leq K\epsilon^{-k}$$

  for some $K > 0$ and $k \geq 1$.

- $\hat{h}_n$ is ϵ-accurate with probability at least $1 - \delta$ if

$$n = \frac{c^2(4C + 2)^2}{2\epsilon^2} \left(\log \left(\frac{2K}{\delta} \right) + k \log \left(\frac{4C + 2}{\epsilon} \right) \right)$$

- Homework: Fix n and δ, derive an upper bound for ϵ.

Remarks

If we want $\epsilon \to 0$ as $n \to \infty$:

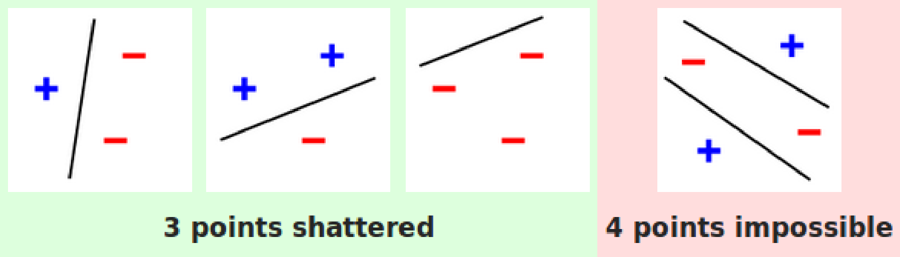
$$dimension(\mathcal{H}) \ll n$$

How do we get that?

- Regularization:
  - Work for the case $dimension(\mathcal{H}) \gg n$
  - Stabilize an estimator $\rightarrow$ force it to live in a neighborhood of a lower-dimensional surface
- Model selection
- Feature selection

Other measures of learning dimension

Vapnik–Chervonenkis dimension



The set of straight lines (as a binary classification model on points) in a two-dimensional plane has VC dimension 3.

Rademacher complexity

- measures richness of a class of real-valued functions *with respect to a probability distribution*
- Given a sample $S = (x_1, x_2, \ldots, x_n)$ and a class $\mathcal{H}$ of real-valued functions defined on the input space $\mathcal{X}$, the empirical Rademacher complexity of $\mathcal{H}$ given S is defined as:

$$Rad(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i) \right]$$

where $\sigma_1, \sigma_2, \ldots, \sigma_m$ are independent random variables drawn from the Rademacher distribution

$$P[\sigma_i = 1] = P[\sigma_i = -1] = 1/2$$

Regularization techniques

Regularization: stability + incorporate special knowledge

Regularization:

- Is superior when the problem is ill-defined, e.g, for the case $\textit{dimension}(\mathcal{H}) \gg n$
- Idea: We want a hypothesis that fits training data well, but also satisfies some good properties
- Stabilize an estimator $\rightarrow$ force it to live in a neighborhood of a lower-dimensional surface

Regularization techniques

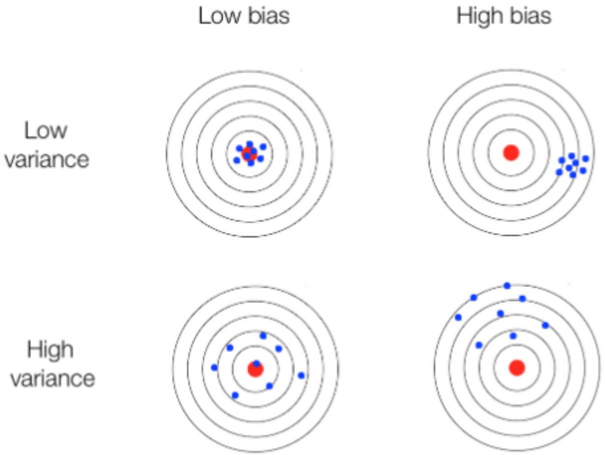
Explicit:

- Adding penalties into loss functions
- Dropout

Implicit:

- Design networks to capture invariance
- Data augmentation
- Use special optimizer
- Early stopping

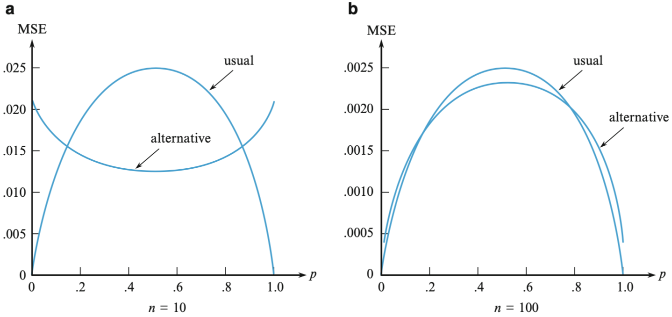
Bias-variance decomposition



$$MSE(\hat{\theta}) = Var(\hat{\theta}) + (bias)^2$$

Stein's phenomenon

- Given i.i.d. $X_1, \ldots, X_n$ samples from $Bernoulli(p)$, we wish to estimate p
- Usual estimate: $(X_1 + X_2 + \ldots X_n)/n$
- Biased estimate: $(X_1 + X_2 + \ldots X_n + 2)/(n + 4)$



Stein's phenomenon: Bernoulli

- Given i.i.d. $X_1, \ldots, X_n$ samples from $\mathcal{N}_d(\mu, I_d)$ $(d \geq 3)$, we wish to estimate μ
- The accuracy of an estimator is measured by the risk function

$$MSE(\hat{\mu}) = E[\|\hat{\mu} - \mu\|^2]$$

- The standard estimate is

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

which minimizes

$$\min_c \sum_{i=1}^{n} \|X_i - c\|^2$$

Stein's phenomenon

- The standard estimate is

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

which minimizes

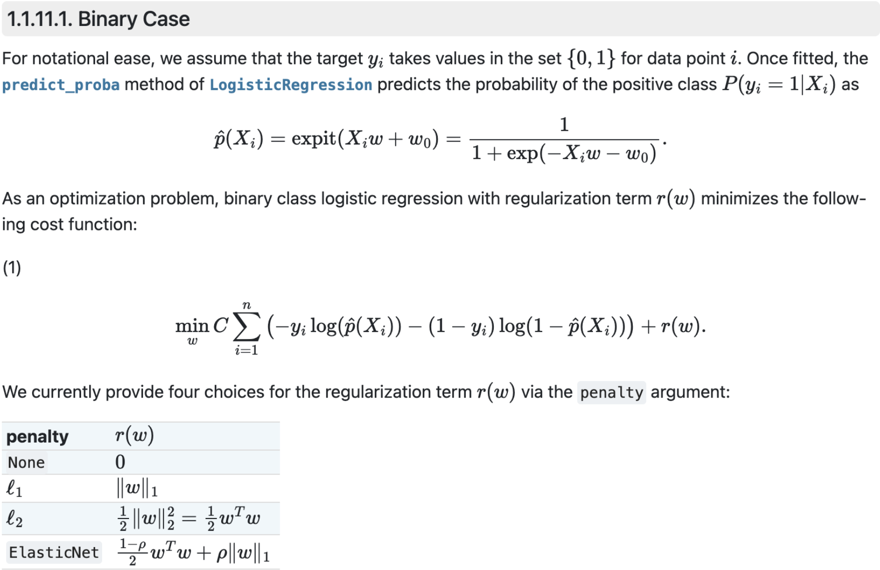
$$\min_c \sum_{i=1}^{n} \|X_i - c\|^2$$

- James-Stein's estimator

$$\mu^{JS} = \left(1 - \frac{d-2}{n\|\bar{X}\|^2}\right)\bar{X}$$

is a strictly better estimator than the sample mean $\bar{X}$

Regularization: LR



1.1.11.1. Binary Case

For notational ease, we assume that the target y_i takes values in the set $\{0, 1\}$ for data point i. Once fitted, the `predict_proba` method of `LogisticRegression` predicts the probability of the positive class $P(y_i = 1|X_i)$ as

$$\hat{p}(X_i) = \operatorname{expit}(X_i w + w_0) = \frac{1}{1 + \exp(-X_i w - w_0)}.$$

As an optimization problem, binary class logistic regression with regularization term $r(w)$ minimizes the following cost function:

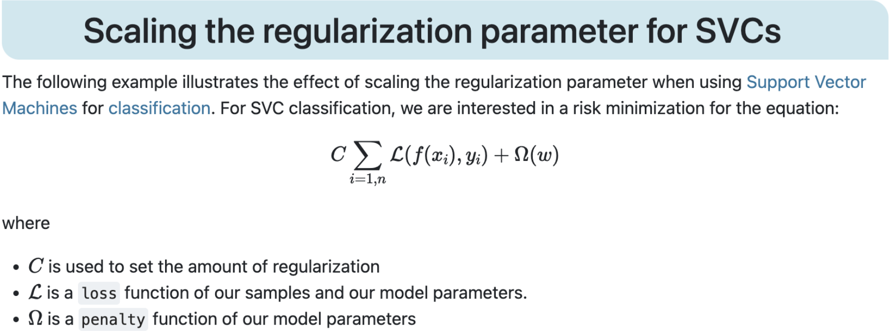
(1)

$$\min_w C \sum_{i=1}^{n} \left(-y_i \log(\hat{p}(X_i)) - (1 - y_i) \log(1 - \hat{p}(X_i))\right) + r(w).$$

We currently provide four choices for the regularization term $r(w)$ via the `penalty` argument:

penalty	$r(w)$
`None`	0
ℓ_1	$\|w\|_1$
ℓ_2	$\frac{1}{2}\|w\|_2^2 = \frac{1}{2}w^T w$
`ElasticNet`	$\frac{1-\rho}{2}w^T w + \rho\|w\|_1$

Regularization: SVM



Scaling the regularization parameter for SVCs

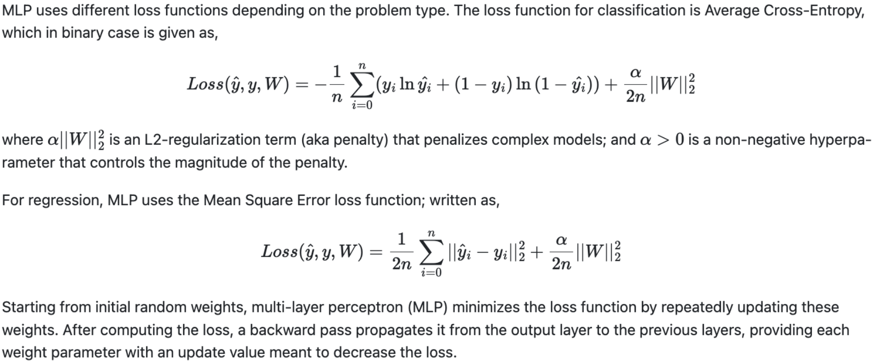
The following example illustrates the effect of scaling the regularization parameter when using Support Vector Machines for classification. For SVC classification, we are interested in a risk minimization for the equation:

$$C \sum_{i=1,n} \mathcal{L}(f(x_i), y_i) + \Omega(w)$$

where

- C is used to set the amount of regularization
- $\mathcal{L}$ is a `loss` function of our samples and our model parameters.
- Ω is a `penalty` function of our model parameters

Regularization: MLP



MLP uses different loss functions depending on the problem type. The loss function for classification is Average Cross-Entropy, which in binary case is given as,

$$Loss(\hat{y}, y, W) = -\frac{1}{n} \sum_{i=0}^{n} (y_i \ln \hat{y}_i + (1 - y_i) \ln (1 - \hat{y}_i)) + \frac{\alpha}{2n} ||W||_2^2$$

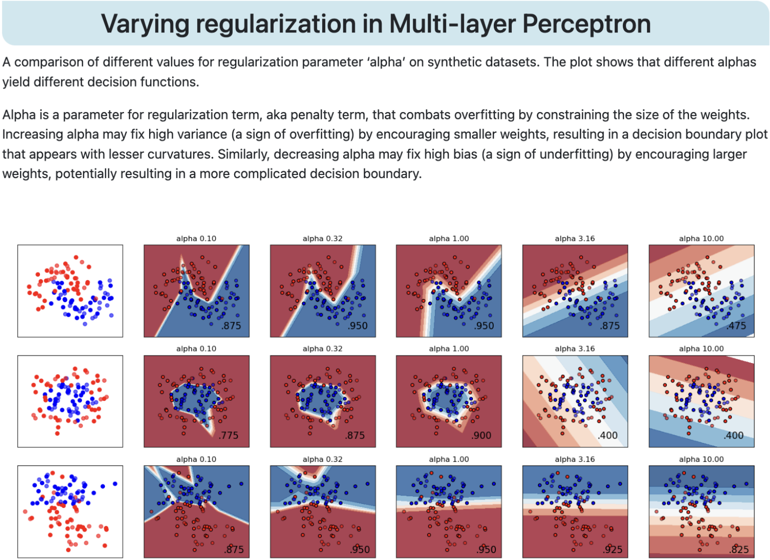
where $\alpha ||W||_2^2$ is an L2-regularization term (aka penalty) that penalizes complex models; and $\alpha > 0$ is a non-negative hyperparameter that controls the magnitude of the penalty.

For regression, MLP uses the Mean Square Error loss function; written as,

$$Loss(\hat{y}, y, W) = \frac{1}{2n} \sum_{i=0}^{n} ||\hat{y}_i - y_i||_2^2 + \frac{\alpha}{2n} ||W||_2^2$$

Starting from initial random weights, multi-layer perceptron (MLP) minimizes the loss function by repeatedly updating these weights. After computing the loss, a backward pass propagates it from the output layer to the previous layers, providing each weight parameter with an update value meant to decrease the loss.

Regularization: MLP

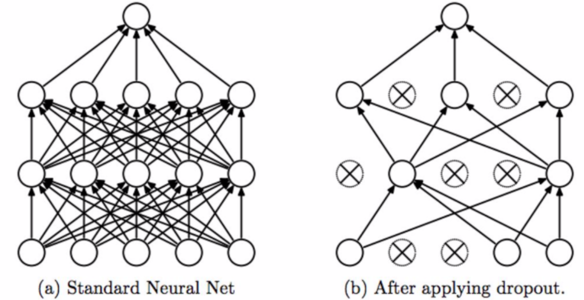


Varying regularization in Multi-layer Perceptron

A comparison of different values for regularization parameter ‘alpha’ on synthetic datasets. The plot shows that different alphas yield different decision functions.

Alpha is a parameter for regularization term, aka penalty term, that combats overfitting by constraining the size of the weights. Increasing alpha may fix high variance (a sign of overfitting) by encouraging smaller weights, resulting in a decision boundary plot that appears with lesser curvatures. Similarly, decreasing alpha may fix high bias (a sign of underfitting) by encouraging larger weights, potentially resulting in a more complicated decision boundary.

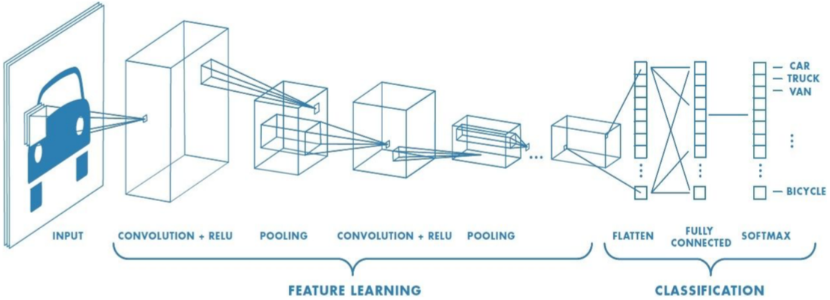
Dropout



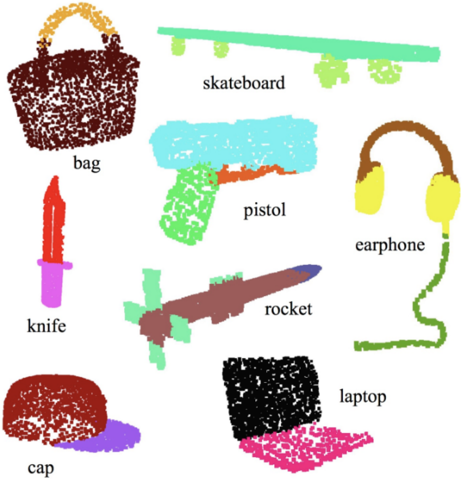
(a) Standard Neural Net

(b) After applying dropout.

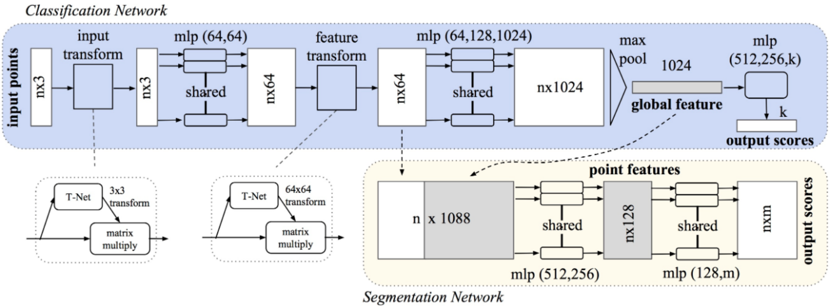
Design networks to capture invariance



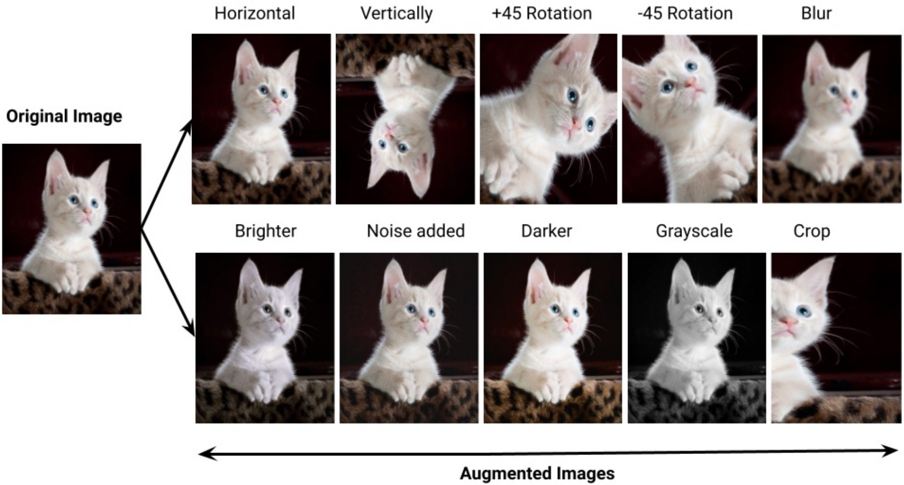
Predictions with point clouds



Design networks to capture invariance

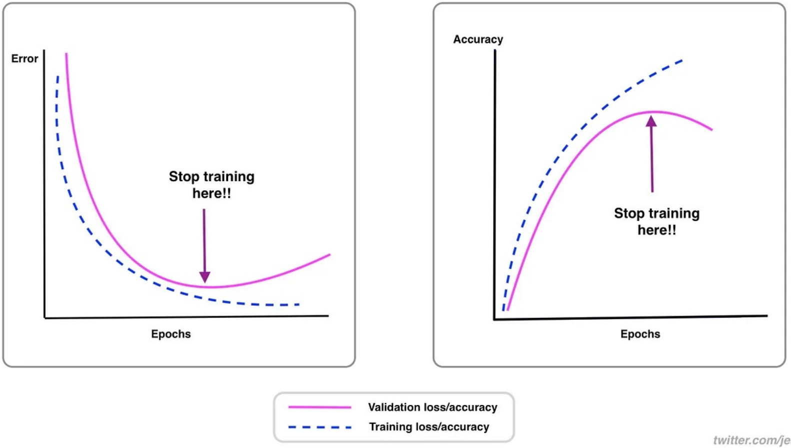


Data augmentation



Early stopping

Early Stopping



twitter.com/jeande_d