

Perceptron and Logistic Regression

Milan Straka

 October 19, 2020



Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



unless otherwise stated

Cross-Validation

We already talked about a **train set** and a **test set**. Given that the main goal of machine learning is to perform well on unseen data, the test set must not be used during training nor hyperparameter selection. Ideally, it is hidden to us altogether.

Therefore, to evaluate a machine learning model (for example to select model architecture, features, or hyperparameter value), we normally need the **validation** or a **development** set.

However, using a single development set might give us noisy results. To obtain less noisy results (i.e., with smaller variance), we can use **cross-validation**.

In cross-validation, we choose multiple validation sets from the training data, and for every one, we train a model on the rest of the training data and evaluate on the chosen validation sets. A commonly used strategy to choose the validation sets is called **k-fold cross-validation**. Here the training set is partitioned into k subsets of approximately the same size, and each subset takes turn to play a role of a validation set.



https://commons.wikimedia.org/wiki/File:K-fold_cross_validation_EN.svg

An extreme case of the **k-fold cross-validation** is **leave-one-out cross-validation**, where every element is considered a separate validation set.

Computing leave-one-out cross-validation is usually extremely inefficient for larger training sets, but in case of linear regression with L2 regularization, it can be evaluated efficiently.

- If you are interested, see:

Ryan M. Rifkin and Ross A. Lippert: Notes on Regularized Least Square

<http://cbcl.mit.edu/publications/ps/MIT-CSAIL-TR-2007-025.pdf>

- Implemented by `sklearn.linear_model.RidgeCV`.

Binary Classification

Binary classification is a classification in two classes.

To extend linear regression to binary classification, we might seek a **threshold** and then classify an input as negative/positive depending whether $y(\mathbf{x}; \mathbf{w}) = \mathbf{x}^T \mathbf{w} + b$ is smaller/larger than a given threshold.

Zero value is usually used as the threshold, both because of symmetry and also because the **bias** parameter acts as a trainable threshold anyway.

Binary Classification

- Consider two points on the decision boundary. Because $y(\mathbf{x}_1; \mathbf{w}) = y(\mathbf{x}_2; \mathbf{w})$, we have $(\mathbf{x}_1 - \mathbf{x}_2)^T \mathbf{w} = 0$, and so \mathbf{w} is orthogonal to every vector on the decision surface – \mathbf{w} is a **normal** of the boundary.
- Consider \mathbf{x} and let \mathbf{x}_\perp be orthogonal projection of \mathbf{x} to the boundary, so we can write $\mathbf{x} = \mathbf{x}_\perp + r \frac{\mathbf{w}}{\|\mathbf{w}\|}$. Multiplying both sides by \mathbf{w}^T and adding b , we get that the distance of \mathbf{x} to the boundary is $r = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}$.
- The distance of the decision boundary from origin is therefore $\frac{|b|}{\|\mathbf{w}\|}$.

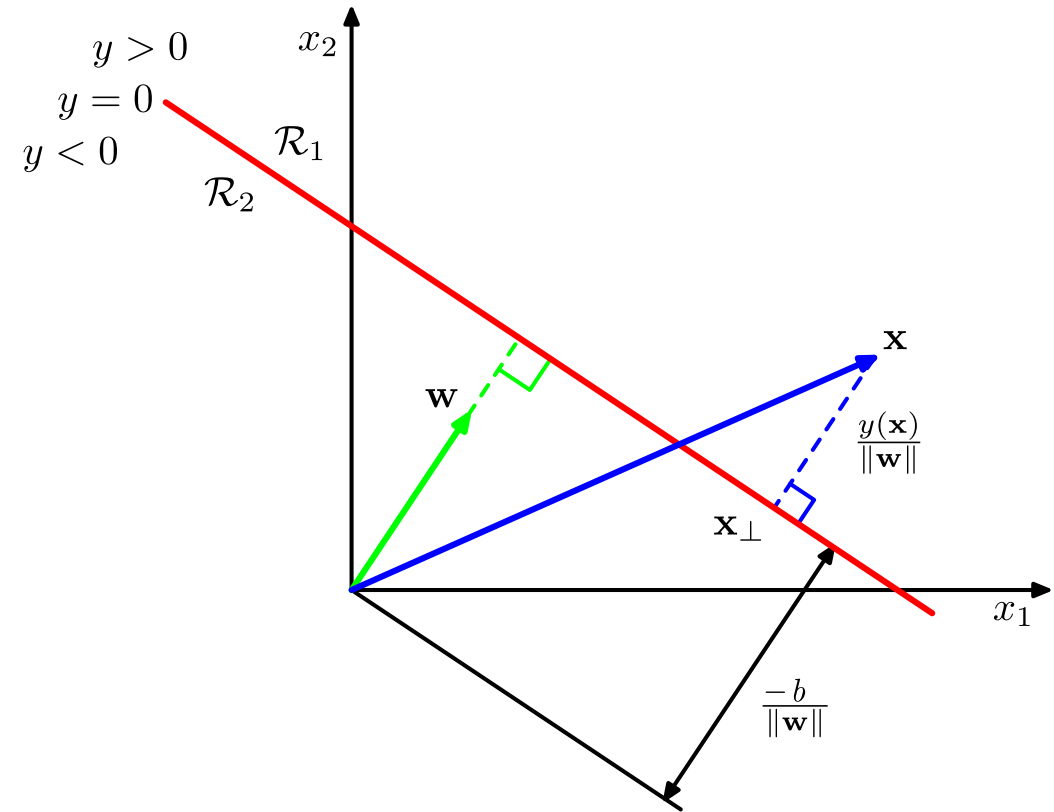


Figure 4.1 of Pattern Recognition and Machine Learning.

The perceptron algorithm is probably the oldest one for training weights of a binary classification. Assuming the target value $t \in \{-1, +1\}$, the goal is to find weights \mathbf{w} such that for all train data,

$$\text{sign}(y(\mathbf{x}_i; \mathbf{w})) = \text{sign}(\mathbf{x}_i^T \mathbf{w}) = t_i,$$

or equivalently,

$$t_i y(\mathbf{x}_i; \mathbf{w}) = t_i \mathbf{x}_i^T \mathbf{w} > 0.$$

Note that a set is called **linearly separable**, if there exists a weight vector \mathbf{w} such that the above equation holds.

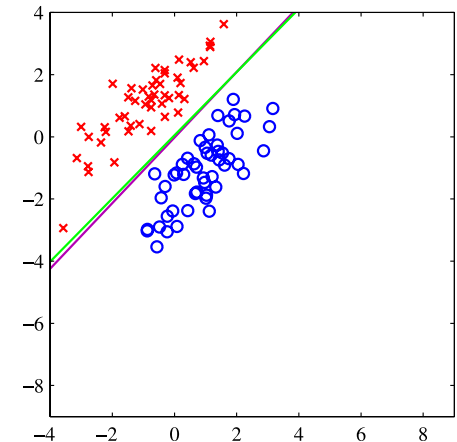


Figure 4.4 of Pattern Recognition and Machine Learning.

The perceptron algorithm was invented by Rosenblatt in 1958.

Input: Linearly separable dataset ($\mathbf{X} \in \mathbb{R}^{N \times D}$, $\mathbf{t} \in \{-1, +1\}^N$).

Output: Weights $\mathbf{w} \in \mathbb{R}^D$ such that $t_i \mathbf{x}_i^T \mathbf{w} > 0$ for all i .

- $\mathbf{w} \leftarrow \mathbf{0}$
- until all examples are classified correctly, process example i :
 - $y \leftarrow \mathbf{x}_i^T \mathbf{w}$
 - if $t_i y \leq 0$ (incorrectly classified example):
 - $\mathbf{w} \leftarrow \mathbf{w} + t_i \mathbf{x}_i$

We will prove that the algorithm always arrives at some correct set of weights \mathbf{w} if the training set is linearly separable.

Consider the main part of the perceptron algorithm:

- $y \leftarrow \mathbf{x}_i^T \mathbf{w}$
- if $t_i y \leq 0$ (incorrectly classified example):
 - $\mathbf{w} \leftarrow \mathbf{w} + t_i \mathbf{x}_i$

We can derive the algorithm using on-line gradient descent, using the following loss function

$$L(y(\mathbf{x}; \mathbf{w}), t) \stackrel{\text{def}}{=} \begin{cases} -t\mathbf{x}^T \mathbf{w} & \text{if } t\mathbf{x}^T \mathbf{w} \leq 0 \\ 0 & \text{otherwise} \end{cases} = \max(0, -t\mathbf{x}^T \mathbf{w}) = \text{ReLU}(-t\mathbf{x}^T \mathbf{w}).$$

In this specific case, the value of the learning rate does not actually matter, because multiplying \mathbf{w} by a constant does not change prediction and does not change the loss derivative. Note that the *second* condition is crucial; the first holds also for logistic regression, but there learning rate matters.

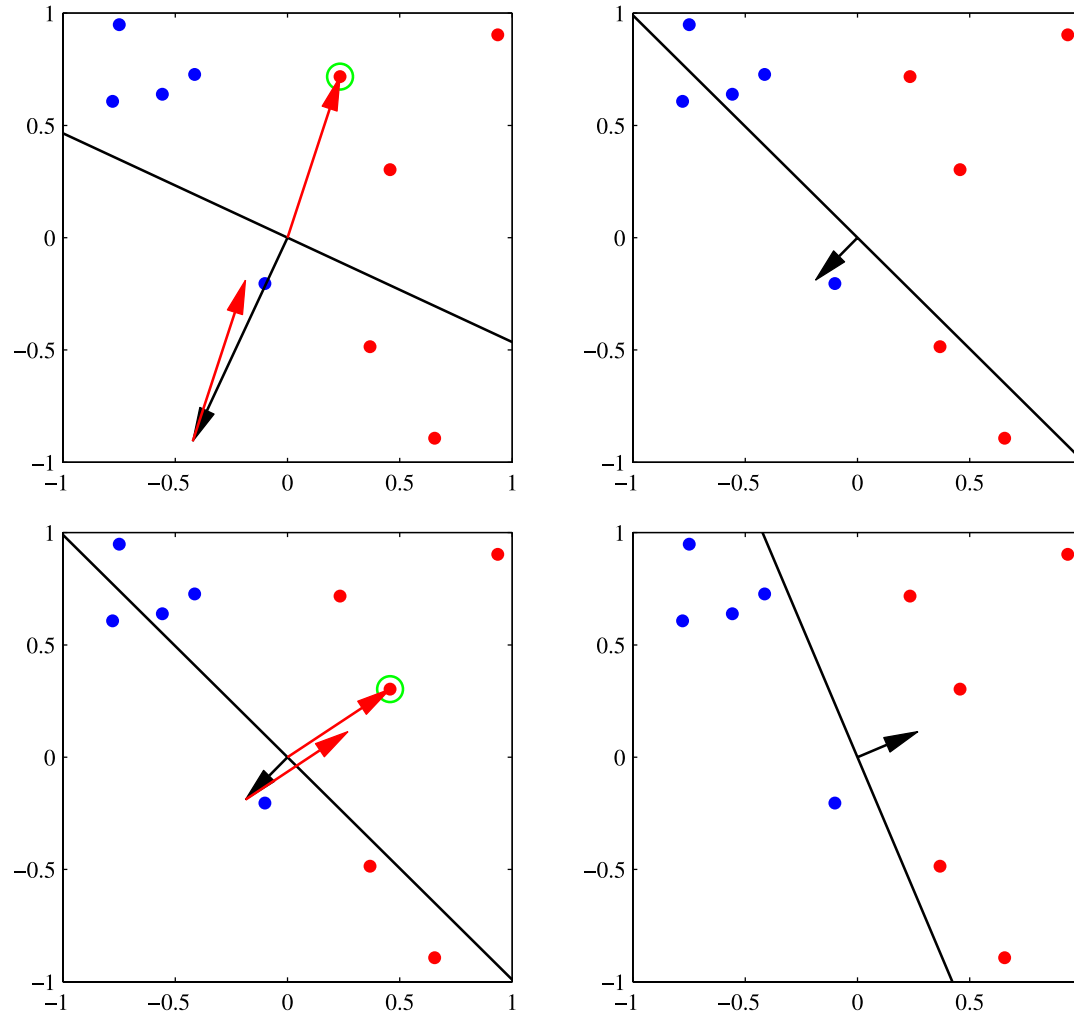


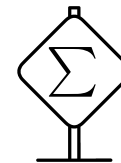
Figure 4.7 of *Pattern Recognition and Machine Learning*.

Proof of Perceptron Convergence

Let \mathbf{w}_* be some weights separating the training data and let \mathbf{w}_k be the weights after k non-trivial updates of the perceptron algorithm, with \mathbf{w}_0 being 0.

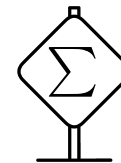
We will prove that the angle α between \mathbf{w}_* and \mathbf{w}_k decreases at each step. Note that

$$\cos(\alpha) = \frac{\mathbf{w}_*^T \mathbf{w}_k}{\|\mathbf{w}_*\| \cdot \|\mathbf{w}_k\|}.$$



Proof of Perceptron Convergence

Assume that the maximum norm of any training example $\|\mathbf{x}\|$ is bounded by R , and that γ is the minimum margin of \mathbf{w}_* , so $t\mathbf{x}^T \mathbf{w}_* \geq \gamma$.



First consider the dot product of \mathbf{w}_* and \mathbf{w}_k :

$$\mathbf{w}_*^T \mathbf{w}_k = \mathbf{w}_*^T (\mathbf{w}_{k-1} + t_k \mathbf{x}_k) \geq \mathbf{w}_*^T \mathbf{w}_{k-1} + \gamma.$$

By iteratively applying this equation, we get

$$\mathbf{w}_*^T \mathbf{w}_k \geq k\gamma.$$

Now consider the length of \mathbf{w}_k :

$$\|\mathbf{w}_k\|^2 = \|\mathbf{w}_{k-1} + t_k \mathbf{x}_k\|^2 = \|\mathbf{w}_{k-1}\|^2 + 2t_k \mathbf{x}_k^T \mathbf{w}_{k-1} + \|\mathbf{x}_k\|^2$$

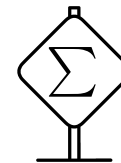
Because \mathbf{x}_k was misclassified, we know that $t_k \mathbf{x}_k^T \mathbf{w}_{k-1} \leq 0$, so $\|\mathbf{w}_k\|^2 \leq \|\mathbf{w}_{k-1}\|^2 + R^2$.

When applied iteratively, we get $\|\mathbf{w}_k\|^2 \leq k \cdot R^2$.

Proof of Perceptron Convergence

Putting everything together, we get

$$\cos(\alpha) = \frac{\mathbf{w}_*^T \mathbf{w}_k}{\|\mathbf{w}_*\| \cdot \|\mathbf{w}_k\|} \geq \frac{k\gamma}{\sqrt{k}R^2\|\mathbf{w}_*\|}.$$



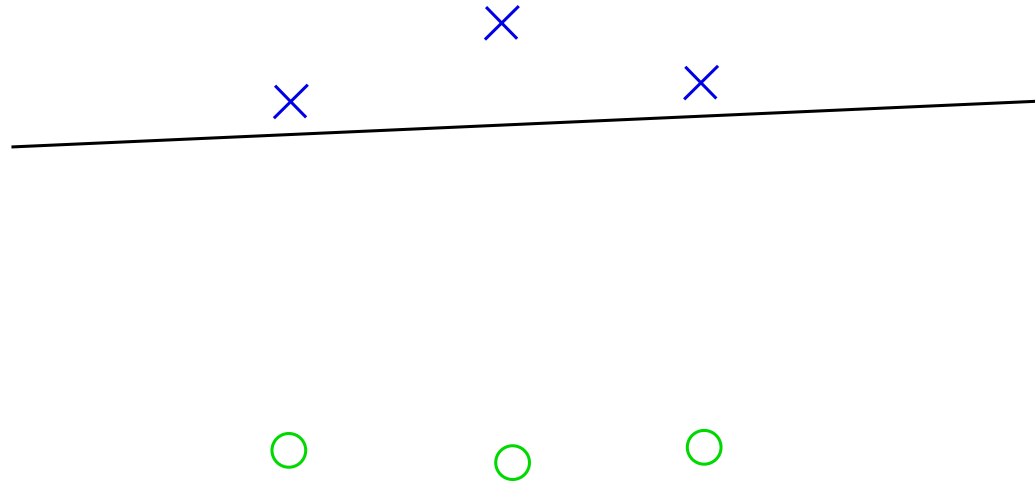
Therefore, the $\cos(\alpha)$ increases during every update. Because the value of $\cos(\alpha)$ is at most one, we can compute the upper bound on the number of steps when the algorithm converges as

$$1 \leq \frac{\sqrt{k}\gamma}{\sqrt{R^2}\|\mathbf{w}_*\|} \text{ or } k \geq \frac{R^2\|\mathbf{w}_*\|^2}{\gamma^2}.$$

Perceptron Issues

Perceptron has several drawbacks:

- If the input set is not linearly separable, the algorithm never finishes.
- The algorithm cannot be easily extended to classification into more than two classes.
- The algorithm performs only prediction, it is not able to return the probabilities of predictions.
- Most importantly, Perceptron algorithm finds *some* solution, not necessary a good one, because once it finds some, it cannot perform any more updates.



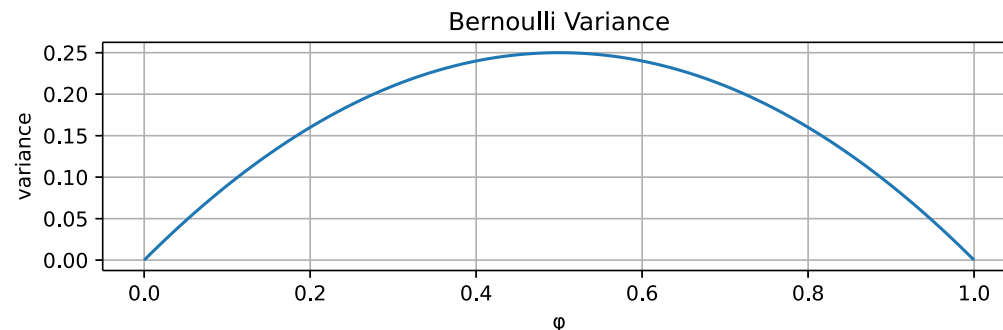
Bernoulli Distribution

The Bernoulli distribution is a distribution over a binary random variable. It has a single parameter $\varphi \in [0, 1]$, which specifies the probability of the random variable being equal to 1.

$$P(x) = \varphi^x (1 - \varphi)^{1-x}$$

$$\mathbb{E}[x] = \varphi$$

$$\text{Var}(x) = \varphi(1 - \varphi)$$



Categorical Distribution

Extension of the Bernoulli distribution to random variables taking one of k different discrete outcomes. It is parametrized by $\mathbf{p} \in [0, 1]^k$ such that $\sum_{i=1}^k p_i = 1$.

$$P(\mathbf{x}) = \prod_i^k p_i^{x_i}$$

$$\mathbb{E}[x_i] = p_i, \text{Var}(x_i) = p_i(1 - p_i)$$

Self Information

Amount of **surprise** when a random variable is sampled.

- Should be zero for events with probability 1.
- Less likely events are more surprising.
- Independent events should have **additive** information.

$$I(x) \stackrel{\text{def}}{=} -\log P(x) = \log \frac{1}{P(x)}$$

Entropy

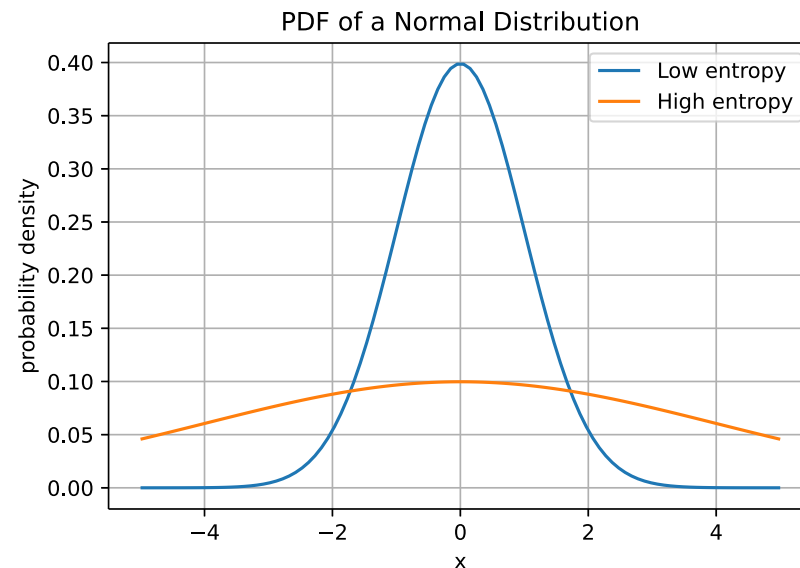
Amount of **surprise** in the whole distribution.

$$H(P) \stackrel{\text{def}}{=} \mathbb{E}_{x \sim P} [I(x)] = -\mathbb{E}_{x \sim P} [\log P(x)]$$

- for discrete P : $H(P) = -\sum_x P(x) \log P(x)$
- for continuous P : $H(P) = -\int P(x) \log P(x) dx$

Note that in the continuous case, the continuous entropy (also called *differential entropy*) has slightly different semantics, for example, it can be negative.

From now on, all logarithms are *natural logarithms* with base e .



Cross-Entropy

$$H(P, Q) \stackrel{\text{def}}{=} -\mathbb{E}_{x \sim P} [\log Q(x)]$$

Gibbs Inequality

- $H(P, Q) \geq H(P)$
- $H(P) = H(P, Q) \Leftrightarrow P = Q$

Proof: Consider $H(P) - H(P, Q) = \sum_x P(x) \log \frac{Q(x)}{P(x)}$.

Using the fact that $\log x \leq (x - 1)$ with equality only for $x = 1$, we get

$$\sum_x P(x) \log \frac{Q(x)}{P(x)} \leq \sum_x P(x) \left(\frac{Q(x)}{P(x)} - 1 \right) = \sum_x Q(x) - \sum_x P(x) = 0.$$

For the equality to hold, $\frac{Q(x)}{P(x)}$ must be 1 for all x , i.e., $P = Q$.

Corollary of the Gibbs inequality

For a categorical distribution with n outcomes, $H(P) \leq \log n$, because for $Q(x) = 1/n$ we get $H(P) \leq H(P, Q) = -\sum_x P(x) \log Q(x) = \log n$.

Nonsymmetry

Note that generally $H(P, Q) \neq H(Q, P)$.

Kullback-Leibler Divergence (KL Divergence)

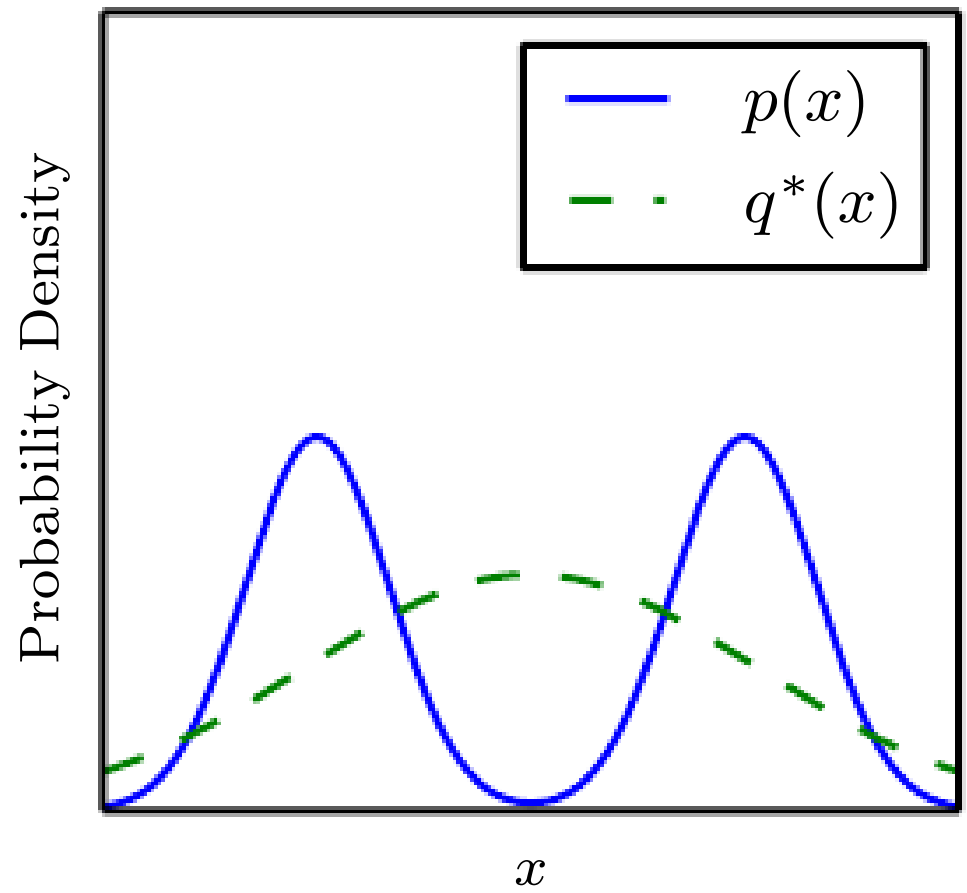
Sometimes also called **relative entropy**.

$$D_{\text{KL}}(P\|Q) \stackrel{\text{def}}{=} H(P, Q) - H(P) = \mathbb{E}_{x \sim P}[\log P(x) - \log Q(x)]$$

- consequence of Gibbs inequality: $D_{\text{KL}}(P\|Q) \geq 0$
- generally $D_{\text{KL}}(P\|Q) \neq D_{\text{KL}}(Q\|P)$

Nonsymmetry of KL Divergence

$$q^* = \operatorname{argmin}_q D_{\text{KL}}(p||q)$$



$$q^* = \operatorname{argmin}_q D_{\text{KL}}(q||p)$$

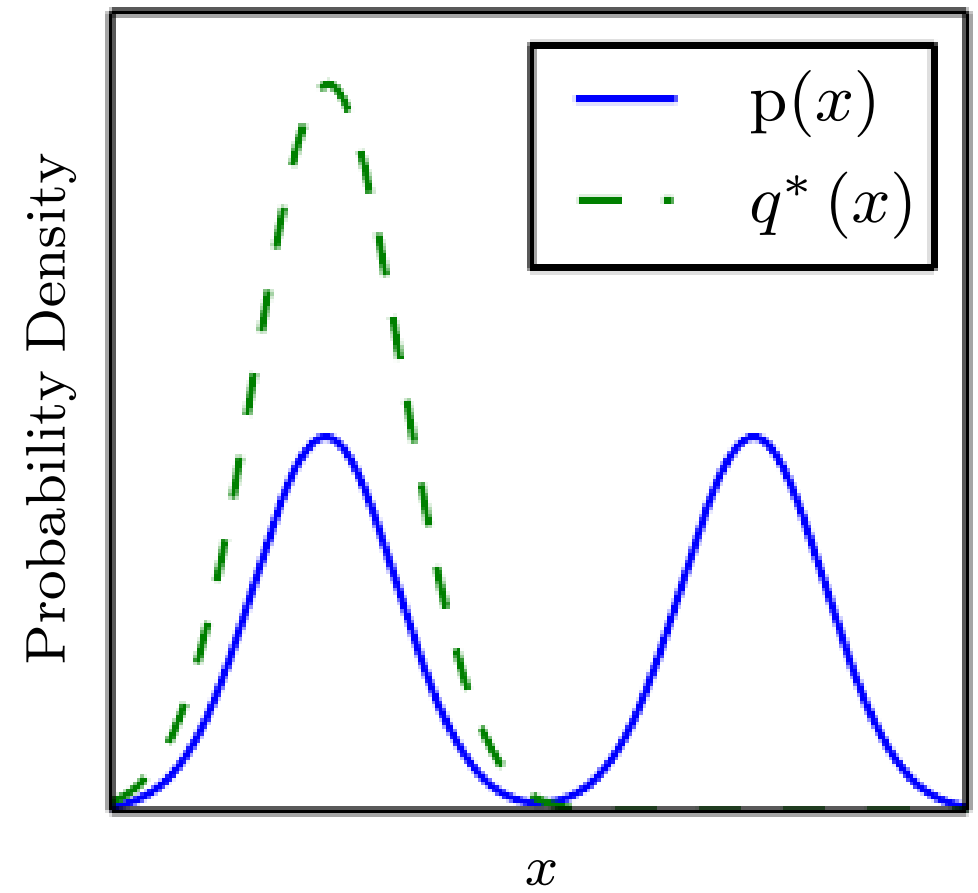


Figure 3.6, page 76 of Deep Learning Book, <http://deeplearningbook.org>

Normal (or Gaussian) Distribution

Distribution over real numbers, parametrized by a mean μ and variance σ^2 :

$$\mathcal{N}(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

For standard values $\mu = 0$ and $\sigma^2 = 1$ we get $\mathcal{N}(x; 0, 1) = \sqrt{\frac{1}{2\pi}} e^{-\frac{x^2}{2}}$.

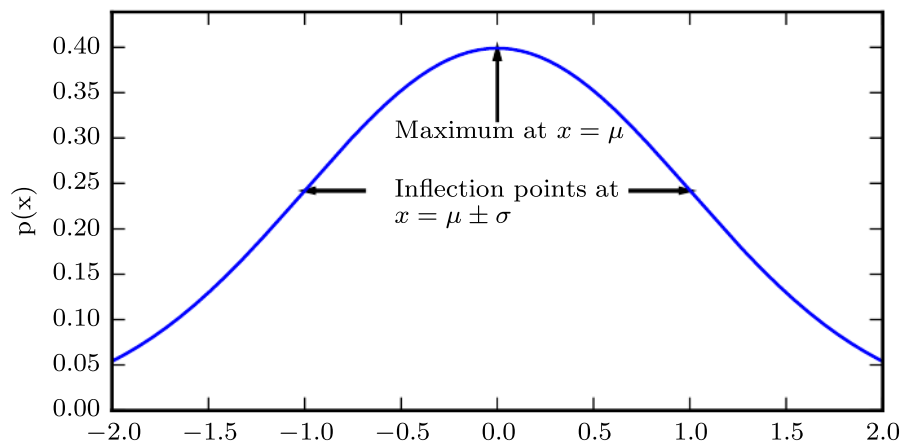


Figure 3.1, page 64 of Deep Learning Book, <http://deeplearningbook.org>.

Central Limit Theorem

The sum of independent identically distributed random variables with finite variance converges to normal distribution.

Principle of Maximum Entropy

Given a set of constraints, a distribution with maximal entropy fulfilling the constraints can be considered the most general one, containing as little additional assumptions as possible.

Considering distributions with a given mean and variance, it can be proven (using variational inference) that such a distribution with **maximum entropy** is exactly the normal distribution.

Maximum Likelihood Estimation

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ be training data drawn independently from the data-generating distribution p_{data} . We denote the empirical data distribution as \hat{p}_{data} .

Let $p_{\text{model}}(\mathbf{x}; \mathbf{w})$ be a family of distributions. The **maximum likelihood estimation** of \mathbf{w} is:

$$\begin{aligned}
 \mathbf{w}_{\text{MLE}} &= \arg \max_{\mathbf{w}} p_{\text{model}}(\mathbf{X}; \mathbf{w}) \\
 &= \arg \max_{\mathbf{w}} \prod_{i=1}^N p_{\text{model}}(\mathbf{x}_i; \mathbf{w}) \\
 &= \arg \min_{\mathbf{w}} \sum_{i=1}^N -\log p_{\text{model}}(\mathbf{x}_i; \mathbf{w}) \\
 &= \arg \min_{\mathbf{w}} \mathbb{E}_{\mathbf{x} \sim \hat{p}_{\text{data}}} [-\log p_{\text{model}}(\mathbf{x}; \mathbf{w})] \\
 &= \arg \min_{\mathbf{w}} H(\hat{p}_{\text{data}}, p_{\text{model}}(\mathbf{x}; \mathbf{w})) \\
 &= \arg \min_{\mathbf{w}} D_{\text{KL}}(\hat{p}_{\text{data}} \| p_{\text{model}}(\mathbf{x}; \mathbf{w})) + H(\hat{p}_{\text{data}})
 \end{aligned}$$

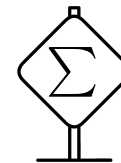
MLE can be easily generalized to a conditional case, where our goal is to predict t given \mathbf{x} :

$$\begin{aligned}\mathbf{w}_{\text{MLE}} &= \arg \max_{\mathbf{w}} p_{\text{model}}(\mathbf{t} | \mathbf{X}; \mathbf{w}) \\ &= \arg \max_{\mathbf{w}} \prod_{i=1}^m p_{\text{model}}(t_i | \mathbf{x}_i; \mathbf{w}) \\ &= \arg \min_{\mathbf{w}} \sum_{i=1}^m -\log p_{\text{model}}(t_i | \mathbf{x}_i; \mathbf{w})\end{aligned}$$

The resulting *loss function* is called *negative log likelihood*, or *cross-entropy* or *Kullback-Leibler divergence*.

Properties of Maximum Likelihood Estimation

Assume that the true data generating distribution p_{data} lies within the model family $p_{\text{model}}(\cdot; \mathbf{w})$. Furthermore, assume there exists a unique $\mathbf{w}_{p_{\text{data}}}$ such that $p_{\text{data}} = p_{\text{model}}(\cdot; \mathbf{w}_{p_{\text{data}}})$.



- MLE is a *consistent* estimator. If we denote \mathbf{w}_m to be the parameters found by MLE for a training set with m examples generated by the data generating distribution, then \mathbf{w}_m converges in probability to $\mathbf{w}_{p_{\text{data}}}$.

Formally, for any $\varepsilon > 0$, $P(\|\mathbf{w}_m - \mathbf{w}_{p_{\text{data}}}\| > \varepsilon) \rightarrow 0$ as $m \rightarrow \infty$.

- MLE is in a sense most *statistically efficient*. For any consistent estimator, we might consider the average distance of \mathbf{w}_m and $\mathbf{w}_{p_{\text{data}}}$, formally $\mathbb{E}_{\mathbf{x}_1, \dots, \mathbf{x}_m \sim p_{\text{data}}} [\|\mathbf{w}_m - \mathbf{w}_{p_{\text{data}}}\|^2]$. It can be shown (Rao 1945, Cramér 1946) that no consistent estimator has lower mean squared error than the maximum likelihood estimator.

Therefore, for reasons of consistency and efficiency, maximum likelihood is often considered the preferred estimator for machine learning.

Logistic Regression

An extension of perceptron, which models the conditional probabilities of $p(C_0|\mathbf{x})$ and of $p(C_1|\mathbf{x})$. Logistic regression can in fact handle also more than two classes, which we will see shortly.

Logistic regression employs the following parametrization of the conditional class probabilities:

$$\begin{aligned}p(C_1|\mathbf{x}) &= \sigma(\mathbf{x}^T \mathbf{w} + b) \\p(C_0|\mathbf{x}) &= 1 - p(C_1|\mathbf{x}),\end{aligned}$$

where σ is a **sigmoid function**

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

It can be trained using the SGD algorithm.

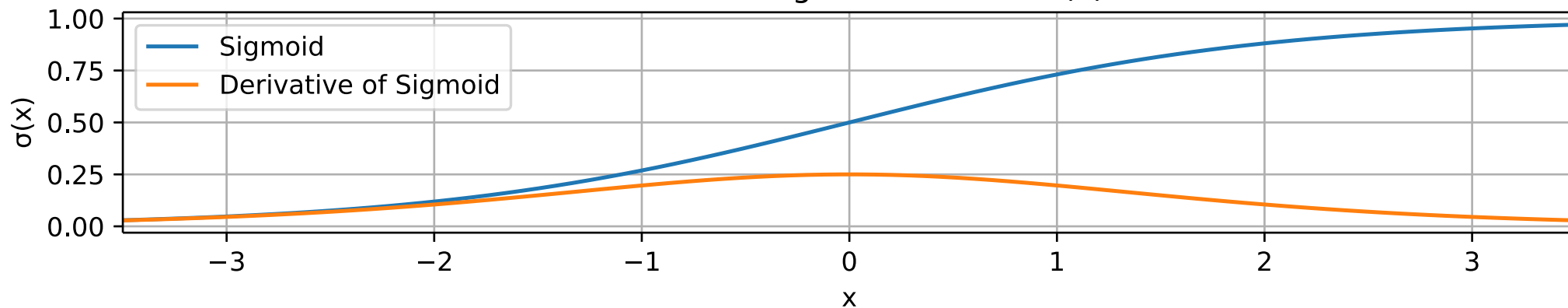
Sigmoid Function

The sigmoid function has values in range $(0, 1)$, is monotonically increasing and it has a derivative of $\frac{1}{4}$ at $x = 0$.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

Plot of the Sigmoid Function $\sigma(x)$



We denote the output of the “linear part” of the logistic regression as

$$\bar{y}(\mathbf{x}; \mathbf{w}) = \mathbf{x}^T \mathbf{w},$$

and the overall prediction as

$$y(\mathbf{x}; \mathbf{w}) = \sigma(\bar{y}(\mathbf{x}; \mathbf{w})) = \sigma(\mathbf{x}^T \mathbf{w}).$$

The logistic regression output $y(\mathbf{x}; \mathbf{w})$ models the probability of class C_1 , $p(C_1|\mathbf{x})$.

To give some meaning to the output of the linear part $\bar{y}(\mathbf{x}; \mathbf{w})$, starting with

$$p(C_1|\mathbf{x}) = \sigma(\bar{y}(\mathbf{x}; \mathbf{w})) = \frac{1}{1 + e^{-\bar{y}(\mathbf{x}; \mathbf{w})}},$$

we arrive at

$$\bar{y}(\mathbf{x}; \mathbf{w}) = \log \left(\frac{p(C_1|\mathbf{x})}{1 - p(C_1|\mathbf{x})} \right) = \log \left(\frac{p(C_1|\mathbf{x})}{p(C_0|\mathbf{x})} \right),$$

which is called a **logit** and it is a logarithm of odds of the probabilities of the two classes.

To train the logistic regression, we use MLE (the maximum likelihood estimation). Its application is straightforward, given that $p(C_1|\mathbf{x}; \mathbf{w})$ is directly the model output $y(\mathbf{x}; \mathbf{w})$.

Therefore, the loss for a batch $\mathbf{X} = \{(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2), \dots, (\mathbf{x}_N, t_N)\}$ is

$$\mathcal{L}(\mathbf{X}) = \frac{1}{N} \sum_i -\log(p(C_{t_i}|\mathbf{x}_i; \mathbf{w})).$$

Input: Input dataset $(\mathbf{X} \in \mathbb{R}^{N \times D}, \mathbf{t} \in \{0, +1\}^N)$, learning rate $\alpha \in \mathbb{R}^+$.

- $\mathbf{w} \leftarrow \mathbf{0}$
- until convergence (or until patience is over), process batch of N examples:
 - $\mathbf{g} \leftarrow \frac{1}{N} \sum_i \nabla_{\mathbf{w}} -\log(p(C_{t_i}|\mathbf{x}_i; \mathbf{w}))$
 - $\mathbf{w} \leftarrow \mathbf{w} - \alpha \mathbf{g}$

Linearity in Logistic Regression

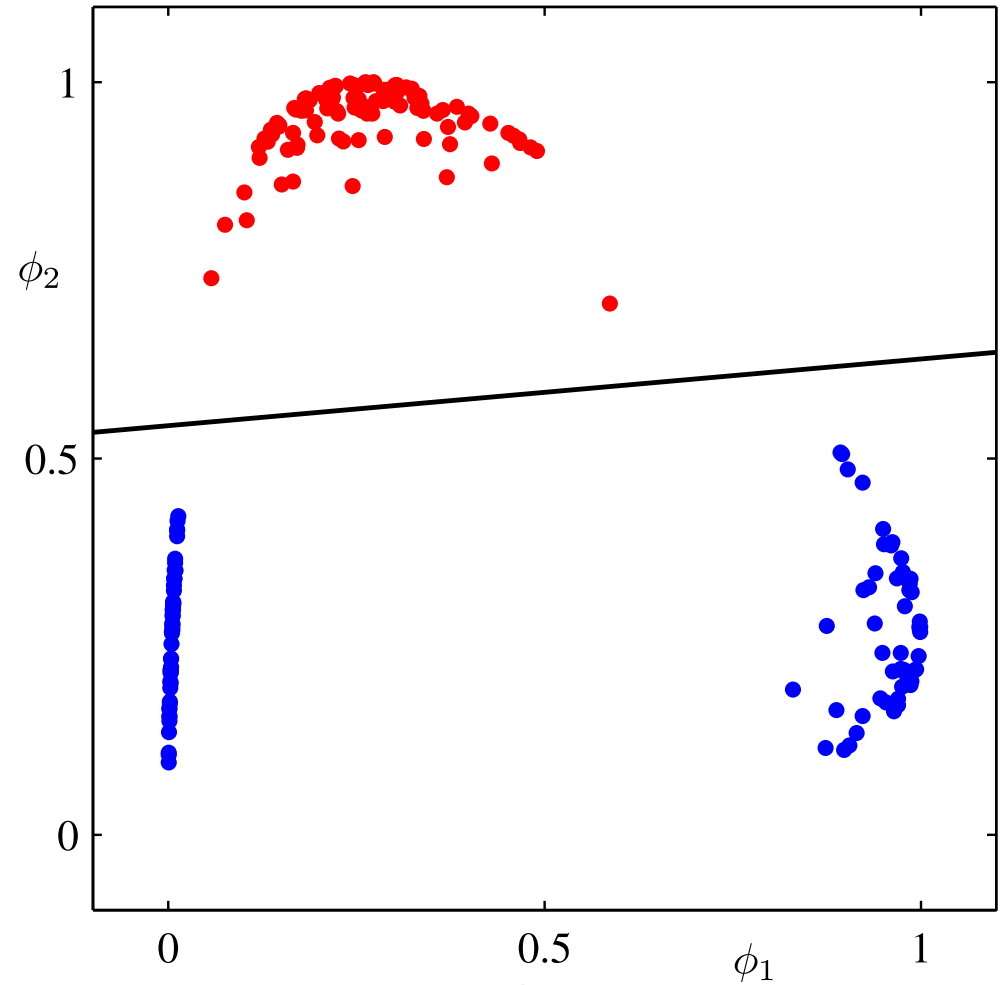
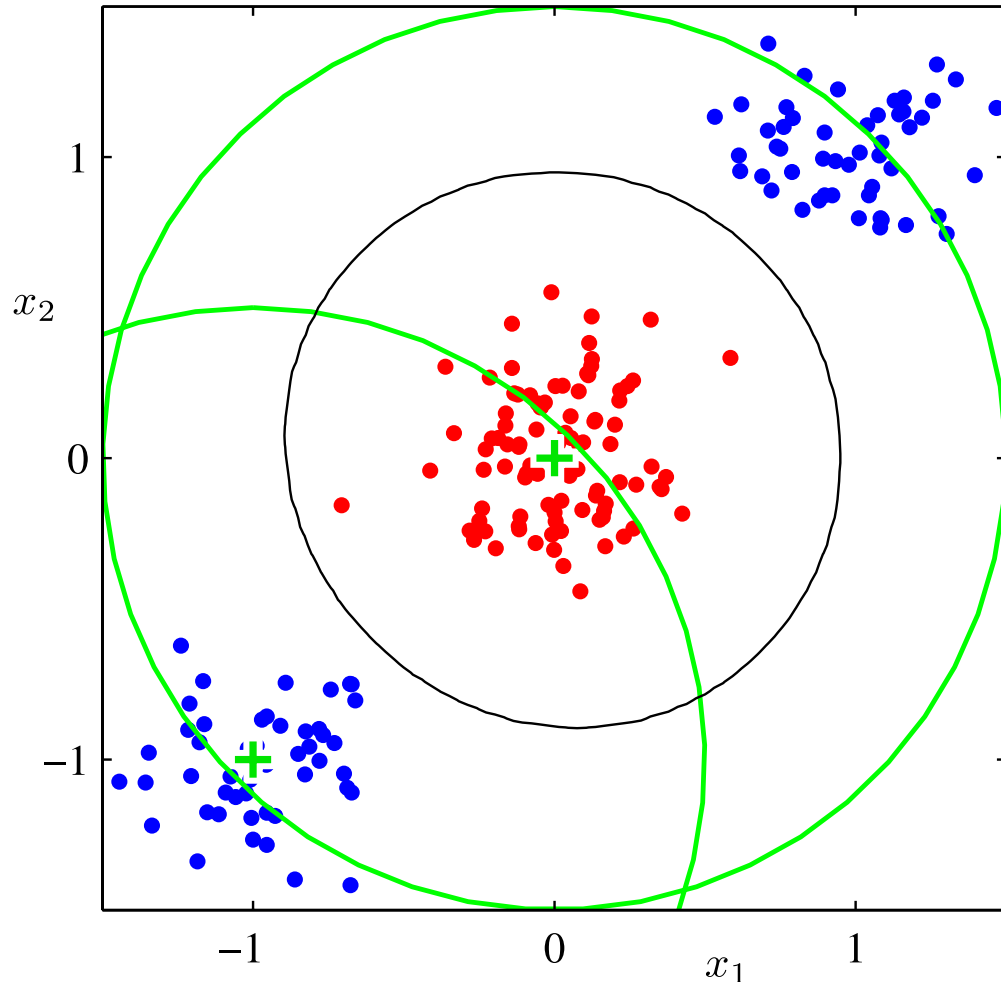


Figure 4.12 of Pattern Recognition and Machine Learning.