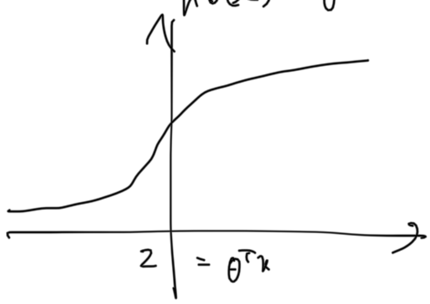
ML w7 Support Vetro Machines

Optimization Objectives.

Logistic Regression

$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$

$h_\theta(x) = g(z)$



If $y=1$, we want $h_\theta(x) \approx 1$, $\theta^T x \gg 0$

If $y=0$, $h_\theta(x) \approx 0$, $\theta^T x \ll 1$

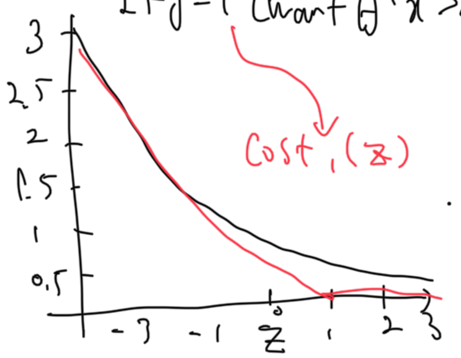
Alternative view of logistic Regression

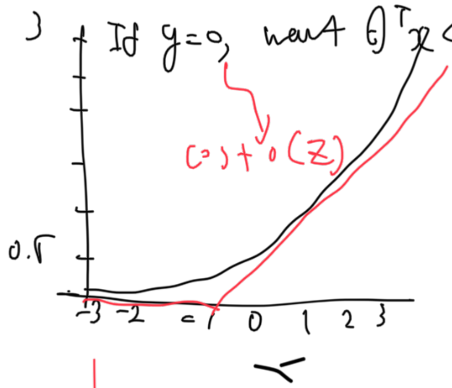
Cost of example: $-(y \log h_\theta(x) + (1-y)\log(1-h_\theta(x)))$

$= -y \log \frac{1}{1+e^{-\theta^T x}} - (1-y)\log(1-\frac{1}{1+e^{-\theta^T x}})$

If $y=1$ (want $\theta^T x \gg 0$)

If $y=0$, want $\theta^T x \ll 0$





Logistic regression

$$\min_\theta \frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \underbrace{(-\log h_\theta(x^{(i)}))}_{cost_1(\theta^T x^{(i)})} + (1-y^{(i)})\underbrace{((-\log(1-h_\theta(x^{(i)})))}_{cost_0(\theta^T x^{(i)})}\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta$$

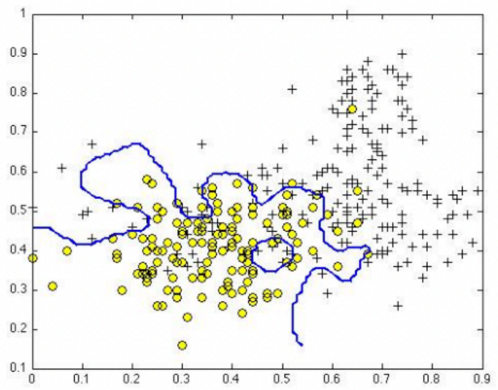
—

Support Vector Machines

最新の提出物の成績評価 40%

1. Suppose you have trained an SVM classifier with a Gaussian kernel, and it learned the following decision boundary on the training set:

0 / 1点



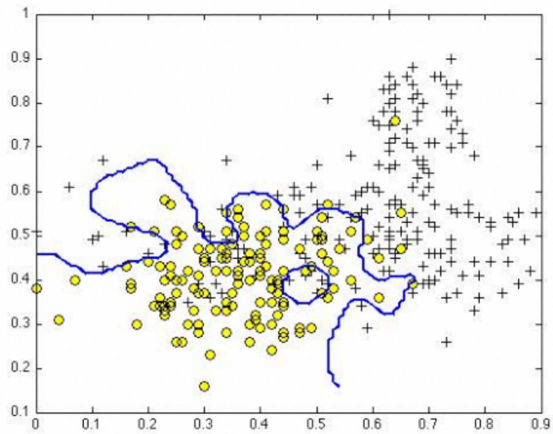
When you measure the SVM's performance on a cross validation set, it does poorly. Should you try increasing or decreasing C? Increasing or decreasing σ^2?

- ○ It would be reasonable to try **decreasing** C. It would also be reasonable to try **increasing** σ^2.
- ○ It would be reasonable to try **decreasing** C. It would also be reasonable to try **decreasing** σ^2.
- ○ It would be reasonable to try **increasing** C. It would also be reasonable to try **increasing** σ^2.
- ◉ It would be reasonable to try **increasing** C. It would also be reasonable to try **decreasing** σ^2.

⊗ 不正解

The figure shows a decision boundary that is overfit to the training set, so we'd like to increase the bias / lower the variance of the SVM. We can do so by either decreasing (not increasing) the parameter C or increasing (not decreasing) σ^2.

1. Suppose you have trained an SVM classifier with a Gaussian kernel, and it learned the following decision boundary on the training set:



When you measure the SVM's performance on a cross validation set, it does poorly. Should you try increasing or decreasing C? Increasing or decreasing σ^2?

- (selected) It would be reasonable to try **decreasing** C. It would also be reasonable to try **increasing** σ^2.
- It would be reasonable to try **decreasing** C. It would also be reasonable to try **decreasing** σ^2.
- It would be reasonable to try **increasing** C. It would also be reasonable to try **decreasing** σ^2.
- It would be reasonable to try **increasing** C. It would also be reasonable to try **increasing** σ^2.

正解

The figure shows a decision boundary that is overfit to the training set, so we'd like to increase the bias / lower the variance of the SVM. We can do so by either decreasing the parameter C or increasing σ^2.

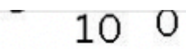
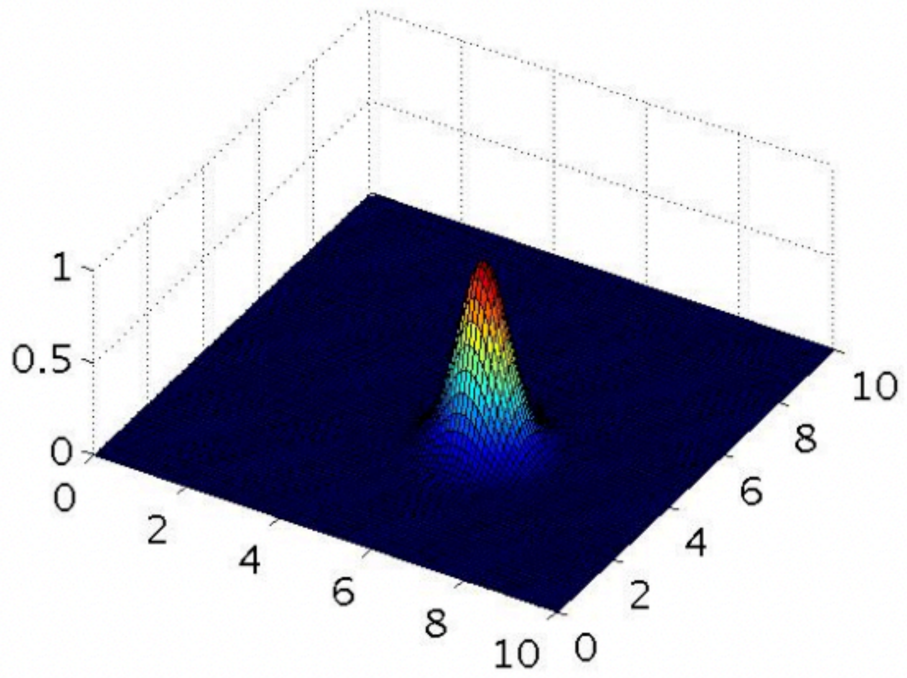


Figure 2.

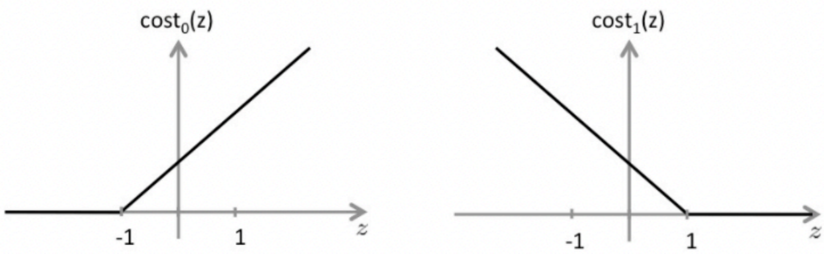


3. The SVM solves

0 / 1点

$$\min_\theta \; C \sum_{i=1}^{m} y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1 - y^{(i)}) \text{cost}_0(\theta^T x^{(i)}) + \sum_{j=1}^{n} \theta_j^2$$

where the functions $\text{cost}_0(z)$ and $\text{cost}_1(z)$ look like this:



The first term in the objective is:

$$C \sum_{i=1}^{m} y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1 - y^{(i)}) \text{cost}_0(\theta^T x^{(i)}).$$

This first term will be zero if two of the following four conditions hold true. Which are the two conditions that would guarantee that this term equals zero?

- [x] For every example with $y^{(i)} = 0$, we have that $\theta^T x^{(i)} \leq -1$.

正解
For examples with $y^{(i)} = 0$, only the $\text{cost}_0(\theta^T x^{(i)})$ term is present. As you can see in the graph, this will be zero for all inputs less than or equal to -1.

- [ ] For every example with $y^{(i)} = 1$, we have that $\theta^T x^{(i)} \geq 0$.
- [ ] For every example with $y^{(i)} = 1$, we have that $\theta^T x^{(i)} \geq 1$.
- [ ] For every example with $y^{(i)} = 0$, we have that $\theta^T x^{(i)} \leq 0$.

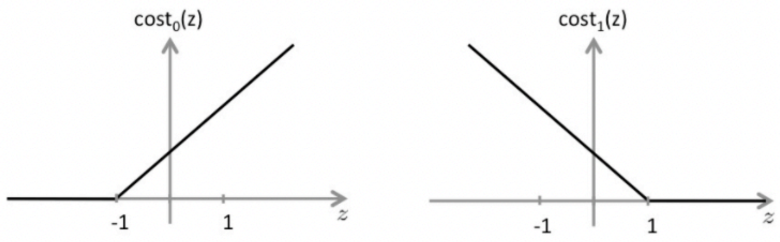
正しい回答をすべて選択しませんでした

3. The SVM solves

0 / 1点

$\min_\theta \; C\sum_{i=1}^{m} y^{(i)}\mathrm{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)})\mathrm{cost}_0(\theta^T x^{(i)}) + \sum_{j=1}^{n}\theta_j^2$

where the functions $\mathrm{cost}_0(z)$ and $\mathrm{cost}_1(z)$ look like this:



The first term in the objective is:

$C\sum_{i=1}^{m} y^{(i)}\mathrm{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)})\mathrm{cost}_0(\theta^T x^{(i)}).$

This first term will be zero if two of the following four conditions hold true. Which are the two conditions that would guarantee that this term equals zero?

- [x] For every example with $y^{(i)} = 0$, we have that $\theta^T x^{(i)} \leq -1$.

正解
For examples with $y^{(i)} = 0$, only the $\mathrm{cost}_0(\theta^T x^{(i)})$ term is present. As you can see in the graph, this will be zero for all inputs less than or equal to -1.

- [x] For every example with $y^{(i)} = 0$, we have that $\theta^T x^{(i)} \leq 0$.

これを選択しないでください
$\mathrm{cost}_0(\theta^T x^{(i)})$ is still non-zero for inputs between -1 and 0, so being less than or equal to 0 is insufficient.

- [ ] For every example with $y^{(i)} = 1$, we have that $\theta^T x^{(i)} \geq 0$.
- [ ] For every example with $y^{(i)} = 1$, we have that $\theta^T x^{(i)} \geq 1$.

4. Suppose you have a dataset with n = 10 features and m = 5000 examples.

0 / 1点

After training your logistic regression classifier with gradient descent, you find that it has underfit the training set and does not achieve the desired performance on the training or cross validation sets.

Which of the following might be promising steps to take? Check all that apply.

- [x] Use an SVM with a Gaussian Kernel.

正解
By using a Gaussian kernel, your model will have greater complexity and can avoid underfitting the data.

- [x] Increase the regularization parameter λ.

これを選択しないでください
You are already underfitting the data, and increasing the regularization parameter only makes underfitting stronger.

- [ ] Use an SVM with a linear kernel, without introducing new features.
- [ ] Create / add new polynomial features.

4. Suppose you have a dataset with n = 10 features and m = 5000 examples.

After training your logistic regression classifier with gradient descent, you find that it has underfit the training set and does not achieve the desired performance on the training or cross validation sets.

Which of the following might be promising steps to take? Check all that apply.

- [ ] Increase the regularization parameter λ.
- [x] Use an SVM with a Gaussian Kernel.

  正解
  By using a Gaussian kernel, your model will have greater complexity and can avoid underfitting the data.

- [x] Create / add new polynomial features.

  正解
  When you add more features, you increase the variance of your model, reducing the chances of underfitting.

- [ ] Use an SVM with a linear kernel, without introducing new features.

5. Which of the following statements are true? Check all that apply.

- [x] The maximum value of the Gaussian kernel (i.e., $sim(x, l^{(1)})$) is 1.

  正解
  When $x = l^{(1)}$, the Gaussian kernel has value $\exp(0) = 1$, and it is less than 1 otherwise.

- [ ] If the data are linearly separable, an SVM using a linear kernel will return the same parameters θ regardless of the chosen value of C (i.e., the resulting value of θ does not depend on C).
- [ ] Suppose you are using SVMs to do multi-class classification and would like to use the one-vs-all approach. If you have K different classes, you will train K - 1 different SVMs.
- [x] It is important to perform feature normalization before using the Gaussian kernel.

  正解
  The similarity measure used by the Gaussian kernel expects that the data lie in approximately the same range.