Lecture 6: Regression problems

We assume that we have $x \in \mathbb{R}^d, y \in \mathbb{R}^d$.

Extend linear classification to (linear) regression, we still have $y^{(i)} = \theta \phi(x^{(i)})$

We will try to find $\theta$ that minimizes $J(\theta) = \sum_{i=1}^{n} (y^{(i)} - \theta \phi(x^{(i)}))^2$

$$\min_j J(\theta) = \min_{\theta} \sum_{i=1}^{n} (y^{(i)} - \theta \phi(x^{(i)}))^2$$

Why is there a problem with minimizing $J(\theta)$? Let’s say I have only one point (pair) in my training set, then I could get many linear boundaries. Which one do I choose? As the dimensionality of the vectors increases the more ill-posed this will become.

Let’s add a regularization term $\frac{\lambda}{2} \| \theta \|^2$ to our sum of the loss $(y^{(i)} - \theta \phi(x^{(i)}) - \theta_0)^2$:

$$J(\theta, \theta_0) = \sum_{i=1}^{n} (y^{(i)} - \theta \phi(x^{(i)}) - \theta_0)^2 + \frac{\lambda}{2} \| \theta \|^2$$

The regularization term will tell us what to choose in the absence of data. We would prefer an answer where $\theta$ is 0.

The effect of the regularization term goes away as you have more examples.

If we drop the offset parameter and assuming $\phi(x) = x$ is the identity mapping, what happens to the line?

**Kernel version**

Let’s add a $\frac{1}{2}$ to the sum: $J(\theta) = \sum_{i=1}^{n} \frac{1}{2} (y^{(i)} - \theta \phi(x^{(i)}))^2 + \frac{\lambda}{2} \| \theta \|^2$

$$\frac{\partial J(\theta)}{\partial \theta} = -\sum_{i=1}^{n} (y^{(i)} - \theta \phi(x^{(i)})) \phi(x^{(i)}) + \lambda \theta = 0$$

Let:

$$\lambda \alpha_i = y^{(i)} - \theta \phi(x^{(i)}) \in \mathbb{R}$$

Then, after substituting in $\frac{\partial J(\theta)}{\partial \theta}$ we have:

$$-\sum_{i=1}^{n} \lambda \alpha_i \phi(x^{(i)}) + \lambda \theta = 0$$

From $\frac{\partial J(\theta)}{\partial \theta} = 0$ we get:
This is called representer’s theorem. Now, we replace the $\theta$ in the definition on $\lambda \alpha$:

$$\lambda \alpha_i = y^{(i)} - \theta \phi(x^{(i)})$$

$$\lambda \alpha_i = y^{(i)} - \theta(\alpha) \phi(x^{(i)}) = y^{(i)} - \sum_{j=1}^{n} \alpha_j \phi(x^{(j)}) \phi(x^{(i)}) = y^{(i)} - \sum_{j=1}^{n} \alpha_j K(x^{(i)}, x^{(j)})$$

$$K(x^{(j)}, x^{(i)}) = \phi(x^{(j)}) \phi(x^{(i)})$$

So we have:

$$\lambda \alpha_i = y^{(i)} - \sum_{j=1}^{n} \alpha_j K(x^{(i)}, x^{(j)})$$

$$\lambda \alpha = y - K \alpha \Rightarrow K \alpha + \lambda \alpha = y \Rightarrow K \alpha + \lambda I \alpha = y \Rightarrow \alpha(K + \lambda I) = y \Rightarrow \alpha = (K + \lambda I)^{-1} y$$

$K$ is positive semi-definite.

Predictions for new point $x$:

$$\hat{y}(x) = \theta(\alpha) \phi(x) = \sum_{i=1}^{n} \alpha_i \phi(x^{(i)}) \phi(x) = \sum_{i=1}^{n} \alpha_i K(x^{(i)}, x) = K_x^T \alpha = K_x^T (K + \lambda I)^{-1} y$$

$$K_x = \begin{bmatrix} K(x^{(1)}, x) \\ \vdots \\ K(x^{(n)}, x) \end{bmatrix}$$

How our solution behaves:

$\lambda \rightarrow \text{very large} \Rightarrow \hat{y}(x) \approx 0$ (less slope allowed on the regression line, see third figure in notebook)

**Model selection**

See figure 4: Which model is correct? Which one will work best with future samples?

- Often cross-validation is very good. Like leave one out cross-validation...

$$\hat{\alpha}_j^{-i} = \text{coefficients computed without } i^{th} \text{ training example}$$

We come up with a predictor:

$$y^{-i}(x) = \sum_{j} \hat{\alpha}_j^{-i} K(x^{(j)}, x)$$

Then we can select the model that minimizes the leave one out cross-validation error:
model = argmin_k \sum_i (y^{(i)} - \hat{y}^k(x^{(i)}))^2

A statistical perspective

\[ y^{(i)} = \theta^* \phi^*(x^{(i)}) + \epsilon_i, \epsilon_i \sim N(0, \sigma^2) \]

\[ y^*(x^{(i)}) = \theta^* \phi^*(x^{(i)}) \]

\[ P(\epsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\epsilon-0)^2} \]

See Figure 5.

\[ E[\hat{y}(x)] - y^*(x) = bias \]

\[ Var \{ \hat{y}(x) \} = variance \]

Complexity of the predictor that I use will impact bias and variance. Inherent bias variance tradeoff.

We will look at models \[ y = \theta \phi(x^{(i)}) + \epsilon, \epsilon \sim N(0, \sigma^2) \] (See figure 6)

\[ E\{y|x\} = \theta \phi(x) \]

\[ Var\{y|x\} = \sigma^2 \]

\[ P(y|x, \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\theta \phi(x))^2} = N(y; \theta \phi(x), \sigma^2) \]

How do we estimate such a model from the data?

**Maximize likelihood:** maximize \( L(\theta, \sigma^2; S_n) = \prod_{i=1}^{n} P(y^{(i)} | x^{(i)}, \theta, \sigma^2) \). This is good when you have a lot of data, since it's an asymptotic approach

**Maximum a posteriori approach:** maximize \( L(\theta, \sigma^2; S_n)P(\theta) \), where likelihood is \( L \) and prior is \( P \).

**Bayesian:** We assume a prior distribution on \( \theta \) and then we compute the posterior distribution:

\[
P(\theta | S_n) = \frac{P(S_n|\theta)P(\theta)}{P(S_n)} = \frac{\int P(S_n|\theta)P(\theta)d\theta}{\int L(\theta; S_n)P(\theta)d\theta} = \frac{1}{z} \frac{P(S_n|\theta)P(\theta)}{P(S_n)} = \frac{1}{z} \frac{P(S_n|\theta)P(\theta)}{P(S_n)}
\]

\[
z = \int L(\theta; S_n)P(\theta)d\theta = \text{marginal likelihood}
\]

Once we adjust the posterior on \( \theta \), we can predict using this posterior probability, instead of the prior:

\[
P(y|x, S_n) = \int_\theta P(y|x, \theta)P(\theta|S_n)d\theta
\]