Gözetimli Öğrenme El Kitabı VIP

Afshine Amidi ve Shervine Amidi

April 30, 2019

Başak Buluz ve Ayyüce Kızrak tarafından çevrilmiştir

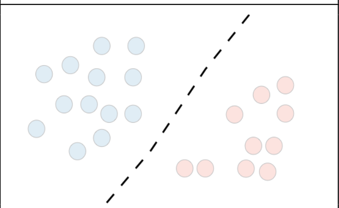
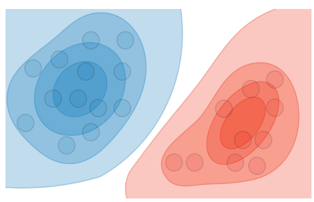
Gözetimli Öğrenmeye Giriş

$\{y^{(1)},...,y^{(m)}\}$ çıktı kümesi ile ilişkili olan $\{x^{(1)},...,x^{(m)}\}$ veri noktalarının kümesi göz önüne alındığında, y'den x'i nasıl tahmin edebileceğimizi öğrenen bir sınıflandırıcı tasarlamak istiyoruz.

❒ **Tahmin türü** – Farklı tahmin modelleri aşağıdaki tabloda özetlenmiştir:

	Regresyon	**Sınıflandırıcı**
Çıktı	Sürekli	Sınıf
Örnekler	Lineer regresyon (bağlanım)	Lojistik regresyon (bağlanım), Destek Vektör Makineleri (DVM), Naive Bayes

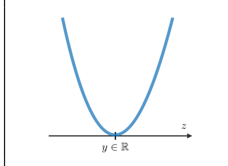
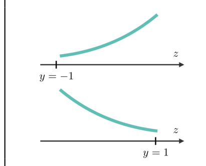
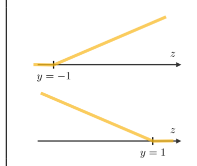
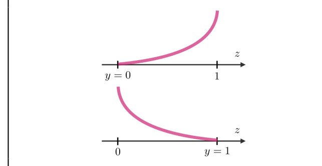
❒ **Model türleri** – Farklı modeller aşağıdaki tabloda özetlenmiştir:

	Ayırt edici model	**Üretici model**
Amaç	Doğrudan tahmin $P(y\|x)$	$P(y\|x)$'i tahmin etmek için $P(x\|y)$'i tahmin etme
Öğrenilenler	Karar Sınırı	Verilerin olasılık dağılımı
Örnekleme		
Örnekler	Regresyon, DVM	GDA, Naive Bayes

Gösterimler ve genel konsept

❒ **Hipotez** – Hipotez h_θ olarak belirtilmiştir ve bu bizim seçtiğimiz modeldir. Verilen $x^{(i)}$ verisi için modelin tahminlediği çıktı $h_\theta(x^{(i)})$'dir.

❒ **Kayıp fonksiyonu** – $L:(z,y)\in\mathbb{R}\times Y\longmapsto L(z,y)\in\mathbb{R}$ şeklinde tanımlanan bir kayıp fonksiyonu y gerçek değerine karşılık geleceği öngörülen z değerini girdi olarak alan ve ne kadar farklı olduklarını gösteren bir fonksiyondur. Yaygın kayıp fonksiyonları aşağıdaki tabloda özetlenmiştir:

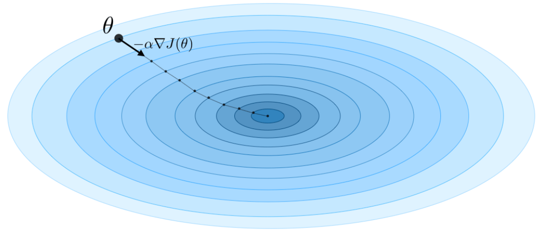
En küçük kareler hatası	**Lojistik yitimi (kaybı)**	**Menteşe yitimi (kaybı)**	**Çapraz entropi**
$\frac{1}{2}(y-z)^2$	$\log(1+\exp(-yz))$	$\max(0,1-yz)$	$-\Big[y\log(z)+(1-y)\log(1-z)\Big]$
			
Lineer regresyon (bağlanım)	Lojistik regresyon (bağlanım)	DVM	Sinir Ağı

❒ **Maliyet fonksiyonu** – J maliyet fonksiyonu genellikle bir modelin performansını değerlendirmek için kullanılır ve L kayıp fonksiyonu aşağıdaki gibi tanımlanır:

$$\boxed{J(\theta)=\sum_{i=1}^m L(h_\theta(x^{(i)}),y^{(i)})}$$

❒ **Bayır inişi** – $\alpha\in\mathbb{R}$ öğrenme oranı olmak üzere, bayır inişi için güncelleme kuralı olarak ifade edilen öğrenme oranı ve J maliyet fonksiyonu aşağıdaki gibi ifade edilir:

$$\boxed{\theta\longleftarrow\theta-\alpha\nabla J(\theta)}$$



Not: Stokastik bayır inişi her eğitim örneğine bağlı olarak parametreyi günceller, ve yığın bayır inişi bir dizi eğitim örneği üzerindedir.

❒ **Olabilirlik** – θ parametreleri verilen bir $L(\theta)$ modelinin olabilirliğini, olabilirliği maksimize ederek en uygun θ parametrelerini bulmak için kullanılır. bulmak için kullanılır. Uygulamada, optimize edilmesi daha kolay olan log-olabilirlik $\ell(\theta)=\log(L(\theta))$'i kullanıyoruz. Sahip olduklarımız:

$$\boxed{\theta^{\textrm{opt}}=\underset{\theta}{\textrm{arg max }}L(\theta)}$$

❒ **Newton'un algoritması** – $\ell'(\theta)=0$ olacak şekilde bir θ bulan nümerik bir yöntemdir. Güncelleme kuralı aşağıdaki gibidir:

$$\boxed{\theta\leftarrow\theta-\frac{\ell'(\theta)}{\ell''(\theta)}}$$

Not: Newton-Raphson yöntemi olarak da bilinen çok boyutlu genelleme aşağıdaki güncelleme kuralına sahiptir:

$$\theta\leftarrow\theta-\left(\nabla_\theta^2\ell(\theta)\right)^{-1}\nabla_\theta\ell(\theta)$$

Lineer regresyon

$y|x;\theta\sim\mathcal{N}(\mu,\sigma^2)$ olduğunu varsayıyoruz

❒ **Normal denklemler** – X matris tasarımı olmak üzere, maliyet fonksiyonunu en aza indiren θ değeri X'in matris tasarımını not ederek, maliyet fonksiyonunu en aza indiren θ değeri kapalı formlu bir çözümdür:

$$\boxed{\theta=(X^TX)^{-1}X^Ty}$$

❒ **En Küçük Ortalama Kareler algoritması** – α öğrenme oranı olmak üzere, m veri noktasını içeren eğitim kümesi için Widrow-Hoff öğrenme oranı olarak bilinen En Küçük Ortalama Kareler Algoritmasının güncelleme kuralı aşağıdaki gibidir:

$$\boxed{\forall j,\quad\theta_j\leftarrow\theta_j+\alpha\sum_{i=1}^m\left[y^{(i)}-h_\theta(x^{(i)})\right]x_j^{(i)}}$$

Not: güncelleme kuralı, bayır yükselişinin özel bir halidir.

❒ **Yerel Ağırlıklı Regresyon** – LWR olarak da bilinen Yerel Ağırlıklı Regresyon ağırlıkları her eğitim örneğini maliyet fonksiyonunda $w^{(i)}(x)$ ile ölçen doğrusal regresyonun bir çeşididir.

$$\boxed{w^{(i)}(x)=\exp\left(-\frac{(x^{(i)}-x)^2}{2\tau^2}\right)}$$

Sınıflandırma ve lojistik regresyon

❒ **Sigmoid fonksiyonu** – Lojistik fonksiyonu olarak da bilinen sigmoid fonksiyonu g, aşağıdaki gibi tanımlanır:

$$\forall z\in\mathbb{R},\quad\boxed{g(z)=\frac{1}{1+e^{-z}}\in]0,1[}$$

❒ **Lojistik regresyon** – $y|x;\theta\sim\textrm{Bernoulli}(\phi)$ olduğunu varsayıyoruz. Aşağıdaki forma sahibiz:

$$\boxed{\phi=p(y=1|x;\theta)=\frac{1}{1+\exp(-\theta^Tx)}=g(\theta^Tx)}$$

Not: Lojistik regresyon durumunda kapalı form çözümü yoktur.

❒ **Softmax regresyonu** – Çok sınıflı lojistik regresyon olarak da adlandırılan Softmax regresyonu 2'den fazla sınıf olduğunda lojistik regresyonu genelleştirmek için kullanılır. Genel kabul olarak, her i sınıfı için Bernoulli parametresi ϕ_i'nin eşit olmasını sağlaması için $\theta_K=0$ olarak ayarlanır.

$$\boxed{\phi_i=\frac{\exp(\theta_i^Tx)}{\displaystyle\sum_{j=1}^K\exp(\theta_j^Tx)}}$$

Genelleştirilmiş Lineer Modeller

❒ **Üstel aile** – Eğer kanonik parametre veya bağlantı fonksiyonu olarak adlandırılan doğal bir parametre η, yeterli bir istatistik $T(y)$ ve aşağıdaki gibi bir log-partition fonksiyonu $a(\eta)$ şeklinde yazılabilirse, dağılım sınıfının üstel ailede olduğu söylenir:

$$\boxed{p(y;\eta)=b(y)\exp(\eta T(y)-a(\eta))}$$

Not: Sık sık $T(y)=y$ olur. Ayrıca, $\exp(-a(\eta))$, olasılıkların birleştiğinden emin olan normalleştirme parametresi olarak görülebilir.

Aşağıdaki tabloda özetlenen en yaygın üstel dağılımlar:

Dağılım	η	$T(y)$	$a(\eta)$	$b(y)$
Bernoulli	$\log\left(\frac{\phi}{1-\phi}\right)$	y	$\log(1+\exp(\eta))$	1
Gauss	μ	y	$\frac{\eta^2}{2}$	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)$
Poisson	$\log(\lambda)$	y	e^η	$\frac{1}{y!}$
Geometrik	$\log(1-\phi)$	y	$\log\left(\frac{e^\eta}{1-e^\eta}\right)$	1

❒ **Genelleştirilmiş Lineer Modellerin Yaklaşımları** – Genelleştirilmiş Lineer Modeller $x\in\mathbb{R}^{n+1}$ için rastgele bir y değişkenini tahminlemeyi hedeflen ve aşağıdaki 3 varsayıma dayanan bir fonksiyondur:

$$(1)\quad\boxed{y|x;\theta\sim\textrm{ExpFamily}(\eta)}\qquad(2)\quad\boxed{h_\theta(x)=E[y|x;\theta]}\qquad(3)\quad\boxed{\eta=\theta^Tx}$$

Not: sıradan en küçük kareler ve lojistik regresyon, genelleştirilmiş doğrusal modellerin özel durumlarıdır.

Destek Vektör Makineleri

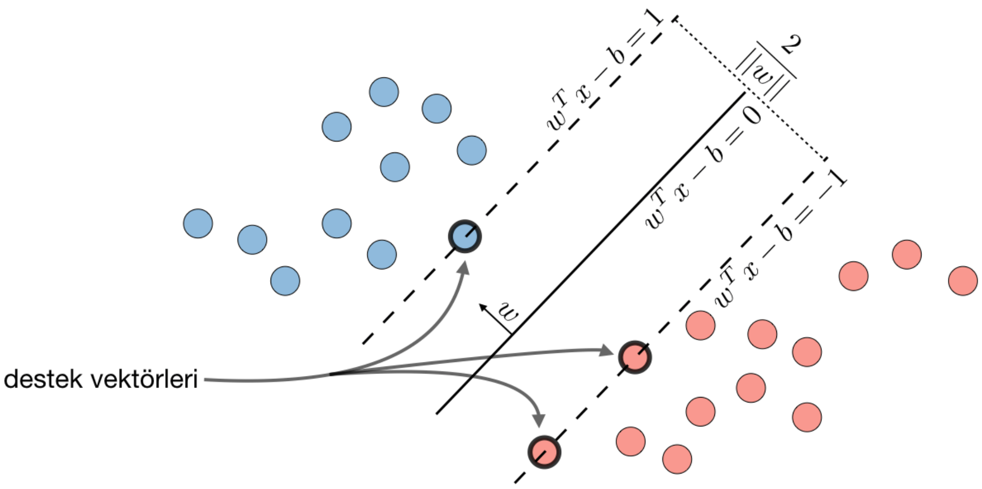
Destek Vektör Makinelerinin amacı minimum mesafeyi maksimuma çıkaran doğruyu bulmaktır.

❒ **Optimal marj sınıflandırıcısı** – h optimal marj sınıflandırıcısı şöyledir:

$$\boxed{h(x)=\textrm{sign}(w^Tx-b)}$$

burada $(w,b)\in\mathbb{R}^n\times\mathbb{R}$, aşağıdaki optimizasyon probleminin çözümüdür:

$$\boxed{\min\frac{1}{2}||w||^2}\quad\textrm{öyle ki}\quad\boxed{y^{(i)}(w^Tx^{(i)}-b)\geqslant1}$$



Not: doğru $\boxed{w^Tx-b=0}$ *şeklinde tanımlanır.*

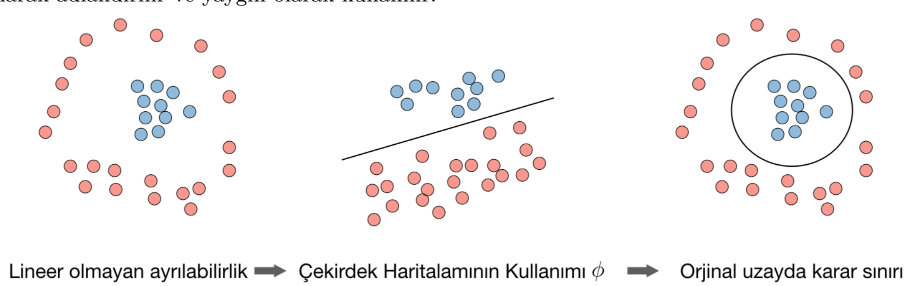
❒ **Menteşe yitimi (kaybı)** – Menteşe yitimi Destek Vektör Makinelerinin ayarlarında kullanılır ve aşağıdaki gibi tanımlanır:

$$\boxed{L(z,y)=[1-yz]_+=\max(0,1-yz)}$$

❒ **Çekirdek** – ϕ gibi bir özellik haritası verildiğinde, K olarak tanımlanacak çekirdeği tanımlarız:

$$\boxed{K(x,z)=\phi(x)^T\phi(z)}$$

Uygulamada, $K(x,z)=\exp\left(-\frac{||x-z||^2}{2\sigma^2}\right)$ tarafından tanımlanan çekirdek K, Gauss çekirdeği olarak adlandırılır ve yaygın olarak kullanılır.



Not: Çekirdeği kullanarak maliyet fonksiyonunu hesaplamak için "çekirdek numarası" nı kullandığımızı söylüyoruz çünkü genellikle çok karmaşık olan ϕ açık haritalamasını bilmeye gerek yok. Bunun yerine, yalnızca $K(x,z)$ değerlerine ihtiyacımız vardır.

❒ **Lagranj** – Lagranj $\mathcal{L}(w,b)$ şeklinde şöyle tanımlanır:

$$\boxed{\mathcal{L}(w,b)=f(w)+\sum_{i=1}^l\beta_ih_i(w)}$$

Not: β_i katsayılarına Lagranj çarpanları denir.

Üretici Öğrenme

Üretken bir model, önce Bayes kuralını kullanarak $P(y|x)$ değerini tahmin etmek için kullanabileceğimiz $P(x|y)$ değerini tahmin ederek verilerin nasıl üretildiğini öğrenmeye çalışır.

Gauss Diskriminant (Ayırtaç) Analizi

❒ **Yöntem** – Gauss Diskriminant Analizi y ve $x|y=0$ ve $x|y=1$ 'in şu şekilde olduğunu varsayar:

$$\boxed{y\sim\textrm{Bernoulli}(\phi)}$$

$$\boxed{x|y=0\sim\mathcal{N}(\mu_0,\Sigma)}\quad\textrm{ve}\quad\boxed{x|y=1\sim\mathcal{N}(\mu_1,\Sigma)}$$

❒ **Tahmin** – Aşağıdaki tablo, olasılığı en üst düzeye çıkarırken bulduğumuz tahminleri özetlemektedir:

$\widehat{\phi}$	$\widehat{\mu_j}\quad(j=0,1)$	$\widehat{\Sigma}$
$\frac{1}{m}\sum_{i=1}^m1_{\{y^{(i)}=1\}}$	$\frac{\sum_{i=1}^m1_{\{y^{(i)}=j\}}x^{(i)}}{\sum_{i=1}^m1_{\{y^{(i)}=j\}}}$	$\frac{1}{m}\sum_{i=1}^m(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T$

Naive Bayes

❒ **Varsayım** – Naive Bayes modeli, her veri noktasının özelliklerinin tamamen bağımsız olduğunu varsayar:

$$\boxed{P(x|y)=P(x_1,x_2,...|y)=P(x_1|y)P(x_2|y)...=\prod_{i=1}^nP(x_i|y)}$$

❒ **Çözümler** – Log-olabilirliğinin $k\in\{0,1\},l\in[\![1,L]\!]$ ile birlikte aşağıdaki çözümlerle maksimize edilmesi:

$$\boxed{P(y=k)=\frac{1}{m}\times\#\{j|y^{(j)}=k\}}\quad\textrm{ve}\quad\boxed{P(x_i=l|y=k)=\frac{\#\{j|y^{(j)}=k\textrm{ ve }x_i^{(j)}=l\}}{\#\{j|y^{(j)}=k\}}}$$

Not: Naive Bayes, metin sınıflandırması ve spam tespitinde yaygın olarak kullanılır.

Ağaç temelli ve topluluk yöntemleri

Bu yöntemler hem regresyon hem de sınıflandırma problemleri için kullanılabilir.

❒ **CART** – Sınıflandırma ve Regresyon Ağaçları (Classification and Regression Trees (CART)), genellikle karar ağaçları olarak bilinir, ikili ağaçlar olarak temsil edilirler.

❒ **Rastgele orman** – Rastgele seçilen özelliklerden oluşan çok sayıda karar ağacı kullanan ağaç tabanlı bir tekniktir. Basit karar ağacının tersine, oldukça yorumlanamaz bir yapıdadır ancak genel olarak iyi performansı onu popüler bir algoritma yapar.

Not: Rastgele ormanlar topluluk yöntemlerindendir.

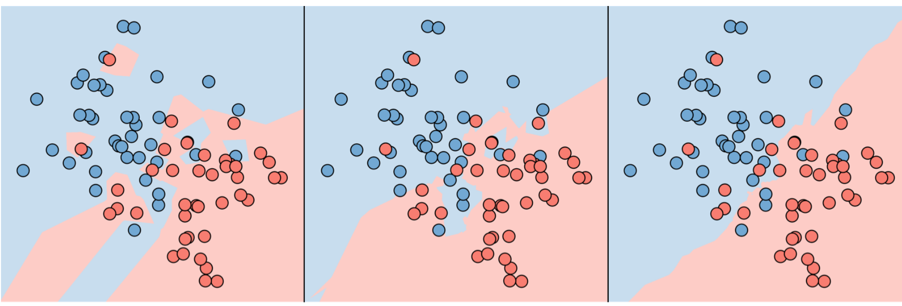
❒ **Artırım** – Artırım yöntemlerinin temel fikri bazı zayıf öğrenicileri biraraya getirerek güçlü bir öğrenici oluşturmaktır. Temel yöntemler aşağıdaki tabloda özetlenmiştir:

Adaptif artırma	**Gradyan artırma**
Yüksek ağırlıklar bir sonraki artırma adımında iyileşmesi için hatalara maruz kalır.	Zayıf öğreniciler kalan hatalar üzerinde eğitildi

Diğer parametrik olmayan yaklaşımlar

❒ ***k*-en yakın komşular** – Genellikle k-NN olarak adlandırılan k-en yakın komşular algoritması, bir veri noktasının tepkisi eğitim kümesindeki kendi k komşularının doğası ile belirlenen parametrik olmayan bir yaklaşımdır. Hem sınıflandırma hem de regresyon yöntemleri için kullanılabilir.

Not: k parametresi ne kadar yüksekse, yanlılık okadar yüksek ve k parametresi ne kadar düşükse, varyans o kadar yüksek olur.

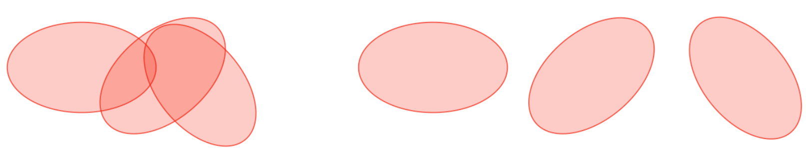


$k = 1$ $k = 3$ $k = 11$

Öğrenme Teorisi

❒ **Birleşim sınırı** – $A_1, ..., A_k$ k olayları olsun. Sahip olduklarımız:

$$\boxed{P(A_1 \cup ... \cup A_k) \leqslant P(A_1) + ... + P(A_k)}$$



$A_1 \cup A_2 \cup A_3$ A_1 A_2 A_3

❒ **Hoeffding eşitsizliği** – $Z_1, .., Z_m$, ϕ parametresinin Bernoulli dağılımından çizilen değişkenler olsun. Örnek ortalamaları mean ve $\gamma > 0$ sabit olsun. Sahip olduklarımız:

$$\boxed{P(|\phi - \widehat{\phi}| > \gamma) \leqslant 2\exp(-2\gamma^2 m)}$$

Not: Bu eşitsizlik, Chernoff sınırı olarak da bilinir.

❒ **Eğitim hatası** – Belirli bir h sınıflandırıcısı için, ampirik risk veya ampirik hata olarak da bilinen eğitim hatasını $\widehat{\epsilon}(h)$ şöyle tanımlarız:

$$\boxed{\widehat{\epsilon}(h) = \frac{1}{m}\sum_{i=1}^{m} 1_{\{h(x^{(i)}) \neq y^{(i)}\}}}$$

❒ **Olası Yaklaşık Doğru** – PAC, öğrenme teorisi üzerine sayısız sonuçların kanıtlandığı ve aşağıdaki varsayımlara sahip olan bir çerçevedir:

- eğitim ve test kümeleri aynı dağılımı takip ediyor
- eğitim örnekleri bağımsız olarak çizilir

❒ **Parçalanma** – $S = \{x^{(1)},...,x^{(d)}\}$ kümesi ve $\mathcal{H}$ sınıflandırıcıların kümesi verildiğinde, $\mathcal{H}$ herhangi bir etiketler kümesi S'e parçalar.

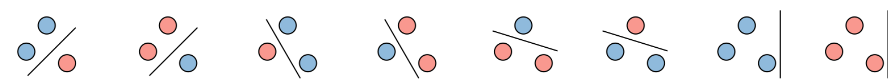
$$\boxed{\exists h \in \mathcal{H}, \quad \forall i \in [\![1,d]\!], \quad h(x^{(i)}) = y^{(i)}}$$

❒ **Üst sınır teoremi** – $|\mathcal{H}| = k$, δ ve örneklem sayısı m'nin sabit olduğu sonlu bir hipotez sınıfı $\mathcal{H}$ olsun. Ardından, en az $1 - \delta$ olasılığı ile elimizde:

$$\boxed{\epsilon(\widehat{h}) \leqslant \left(\min_{h \in \mathcal{H}} \epsilon(h)\right) + 2\sqrt{\frac{1}{2m}\log\left(\frac{2k}{\delta}\right)}}$$

❒ **VC boyutu** – VC($\mathcal{H}$) olarak ifade edilen belirli bir sonsuz $\mathcal{H}$ hipotez sınıfının Vapnik-Chervonenkis (VC) boyutu, $\mathcal{H}$ tarafından parçalanan en büyük kümenin boyutudur.

Not: $\mathcal{H} = \{$2 boyutta doğrusal sınıflandırıcılar kümesi$\}$'nin VC boyutu 3'tür.



❒ **Teorem (Vapnik)** – $\mathcal{H}$, VC($\mathcal{H}$) $= d$ ve eğitim örneği sayısı m verilmiş olsun. En az $1 - \delta$ olasılığı ile, sahip olduklarımız:

$$\boxed{\epsilon(\widehat{h}) \leqslant \left(\min_{h \in \mathcal{H}} \epsilon(h)\right) + O\left(\sqrt{\frac{d}{m}\log\left(\frac{m}{d}\right) + \frac{1}{m}\log\left(\frac{1}{\delta}\right)}\right)}$$